



Seminar Datenqualität & Seminar Optimierungs- und  
Modellierungstechniken für Datenbanken  
Einführung  
21. Oktober 2009

---

# Agenda

## Allgemeines

### Themen

- DQ
- Optimierung
- Schnittstellenthemen

### Seminardurchführung

### Themenvergabe

---

# Allgemeines: Seminarablauf

## Leistungen:

- 1 Vortrag
- Ausarbeitung (ca. 12 Seiten pro Teilnehmer davon 8 Inhalt)

## 4 Termine

- Einführungsveranstaltung am 21.10.2009 (heute)
- Präsentationen am 03./10./17.12.2010 (ab 17 Uhr)
- Abgabe der Ausarbeitung 01.02.2010

## Bewertung

- Vorträge ~40–50% (20 min Vortrag+5–10 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

Unterschiedliche  
Repräsentationen

Widersprüchliche  
Werte

Referentielle Integrität  
verletzt

unvollständig

Tab.: Person

KID	KName	Gebdat	Alter	Geschlecht	Telefon	PLZ	Email
34	Meier, Tom	21.01.1980	35	M	999-999	10117	null
34	Tina Möller	18.04.78	29	W	763-222	36999	null
35	Tom Meier	32.05.1969	27	F	222-231	10117	t@r.de

Eindeutigkeit verletzt

Tab.: Ort

PLZ	Ort
10117	Berlin
36996	Spanien
95555	Ullm

Fehlende Werte  
(z.B.: Default-  
Wert)

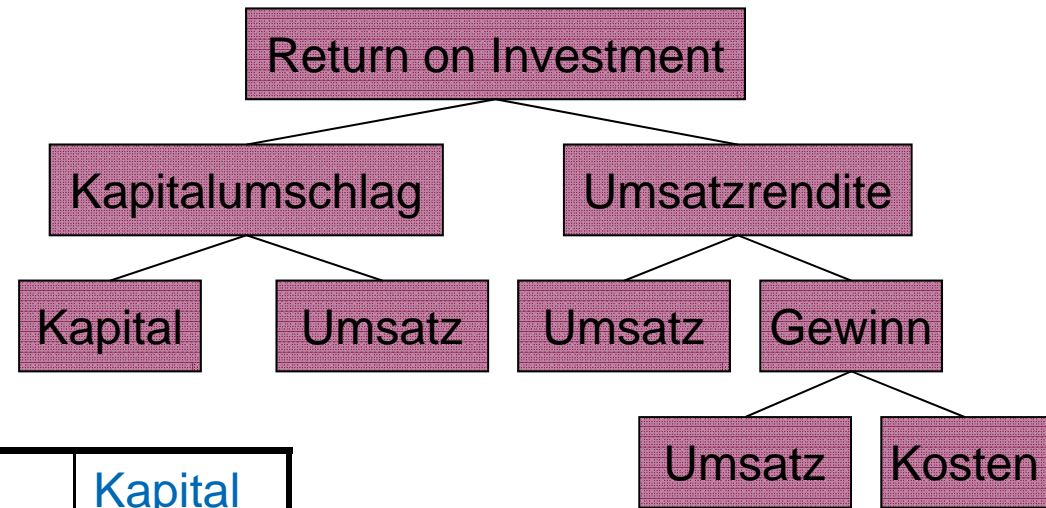
Duplikate

Falsche oder  
unzulässige Werte

Schreib- oder  
Tippfehler

# 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

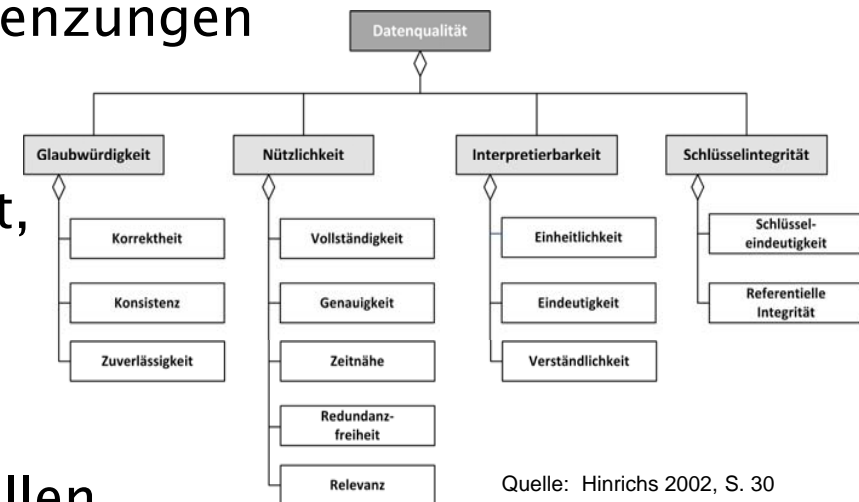
# Themengebiete: DQ im Allgemeinen

## Überblick Datenqualität

- Begriffe Daten und Qualität, Abgrenzungen

## Dimensionen von Datenqualität

- z.B. Glaubwürdigkeit, Nützlichkeit, Interpretierbarkeit, Integrität



## Überblick über Datenqualitätsmodellen

- Welche Modelle gibt es für DQ

## Vorgehensmodell für Datenqualität

- Wie wird Datenqualität erzielt?
- Einsatz von Tools (Toolvorgaben)

---

## Themengebiete: DQ – Spezialthemen

### Data Edits

- Fellegi Holt (1976) (Statistical Edits), viele Erweiterungen
- Numerische Zusammenhänge

### Objekt-Identifikation

- Beispiel: Mietinserat

### Record-Linkage

- Verknüpfung von Datensätzen bei fehlendem PK

### Tools für DQ

- Überblick über DQ-Abfragen und ihre Herausforderungen

# Themengebiete: Optimierung

## Streaming DWH

- Welche Ansätze gibt es überhaupt
- Was sind Besonderheiten im Gegensatz zu "normalen" DWH?
- Szenarien/Einsatzgebiete und/oder Ansätze mit eingebetteten Systemen als Quellen

## Security und DBMS

- Was sind Sicherheitsmechanismen in aktuellen DBS und DWH?
- Welche Sicherheitsaspekte werden durch den jeweiligen Mechanismus abgedeckt
- Was sind eigentlich aktuell die größten Gefahren für DBS/DWH?
- Big Picture – Abstrakte Zusammenfassung, der Mechanismus XYZ wird häufig eingesetzt um ... zu schützen

---

# Themengebiete: Optimierung

## Stromverbrauch in DBMS/Green IT

### State of the art

- Überblick zu aktueller Entwicklung im Bereich (Software)
- Zusammenfassung der derzeitigen Entwicklung
- Kritische Diskussion der verschiedenen Ansätze

### Spezielle DBMS für Stromverbrauchsminimierung

- Welche Ansätze gibt es in diesem Bereich
- Gibt es spezielle alternative Algorithmen um den Stromverbrauch zu senken (Indexe, Sortierung, Joins, ...)?

### Messung und/oder Berechnung

- Gibt es Verfahren zur Messung oder Berechnung des Stromverbrauches von Software?
- State of the art und kritische Diskussion der Ansätze

# Themengebiete: Optimierung

## DBMS in Sensornetzwerken

- Übersicht über bestehende DBMS für Sensornetzwerke
- Worin unterscheiden sich die einzelnen DBMS (Funktion), und wo werden diese eingesetzt?
- Unterschiede zu "normalen" DBMS → Welche Zusatzfunktionen werden benötigt oder können weggelassen werden?

## Real-time DWH

- Wo setze die Ansätze an → nur physische Ebene oder Einfluss auf höhere Ebenen?
- Was bedeutet real-time aktuell in Bezug auf DWH?
- Gibt es Messungen zu diesen Ansätzen?
- Gibt es zero-latency Ansätze/Systeme?

# Themengebiete: Optimierung

## Maßschneiderung von DBMS

### Möglichkeiten der Maßschneiderung

- Was kann alles maßgeschneidert werden? (Bsp. SQL à la carte für SQL Dialekte)
- In welchen Bereichen gib es schon Maßschneiderung und welche Techniken werden dafür verwendet?
- Vergleich und Diskussion verschiedener Ansätze

### Maßgeschneiderte Datenhaltung

- Welche Ansätze gibt es?
- Worin liegen die Unterschiede/das Potenzial?

### Maßgeschneiderte Anfrageoptimierung

- Gibt es aktuelle Ansätze? Wie unterscheiden sich diese?
- Physisch vs. funktional

---

# Themengebiete: Optimierung

## Hybride DBS

- Welche Art von hybriden DBS gibt es?
- Unter welchem Aspekt wurden/werden sie entwickelt?
- Welche Relevanz haben diese Systeme?
- Diskussion der verschiedenen Ansätze

## Column Stores

### Query Processing in Column Stores

- State of the art
- Unterschiede und Diskussion der Ansätze/Anwendungsszenarien

### Technische Aspekte

- Unterschiedliche Techniken bei der Umsetzung von Column Stores
- Benchmarking/Vergleichbarkeit der Ansätze

## Themengebiete: Schnittstelle

Workload bei Datenqualitätsaufgaben

- DQ relativ aufwendig, welche Strategien sind möglich

Optimierung der DQ–Aufgaben innerhalb einer Relation

- Spaltenanalyse, Abhängigkeitsanalyse

Optimierung der DQ–Aufgaben Relationen–übergreifend

- Beziehungsanalyse, Regelbasierte Analyse

Optimierung und Potentiale im DWH–Kontext

- OLAP, Cubes, multidimensionale

---

## Vortrag

- 20 Minuten Vortrag
- 5–10 Minuten Diskussion/Fragen
  
- Überziehen: Redner wird abgewürgt
- Zu früh: Mehr Fragen (ggf. mehr Kritik)
  
- Notebook 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-serif 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 = Ultimative 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:

- Originaler 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 12) überschreiten, aber auch nicht nur Hälfte nutzen
- Nicht an Vorlage halten

# Themenvergabe

Thema	Bearbeiter	Thema	Bearbeiter
DQ-Überblick	Braun	DBMS in Sensors	Maekeler
DQ-Dimensionen	Ahmet	Hybride DBMS	Stegen
DQ-Modelle	Hart	CS QueryProcess	Wolfslast
Data-Edits	Marth	Security & DBMS	Polifka
Objekt-Ident	Hielscher	DBMS min Strom	Stephan
Tools	Tobis	Maß. Anfrageoptim.	Beyer
Workload	Röhl	Green IT SotA	Clausing
RT DWH	Wicht	Möglichkeiten d. Maß.	Warschewske
Streaming DWH	Lux	Maß. Datenhaltung	Lan Si

Danke für die Aufmerksamkeit

[www.witi.cs.uni-magdeburg.de/iti\\_db](http://www.witi.cs.uni-magdeburg.de/iti_db)

Dr. V. Köppen: [vkoeppen@ovgu.de](mailto:vkoeppen@ovgu.de)

A. Lübcke: [luebcke@ovgu.de](mailto:luebcke@ovgu.de)

[www.ovgu.de](http://www.ovgu.de)

---

## Literatur

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