30. Juni 2006 - Technische Universität Kaiserslautern

Grundlagen der Datenintegration

Mastering the Information Explosion -
Information Integration and Information Quality

Paul R. Schilling

Agenda

1. Motivation
2. Die Informationsintegration als Teil der Unternehmensintegration
3. Theorie der Datenintegration - Definitionen
4. Mechanismen der Datenintegration - Replikationsorientierte vs. Virtuelle Integration

Agenda

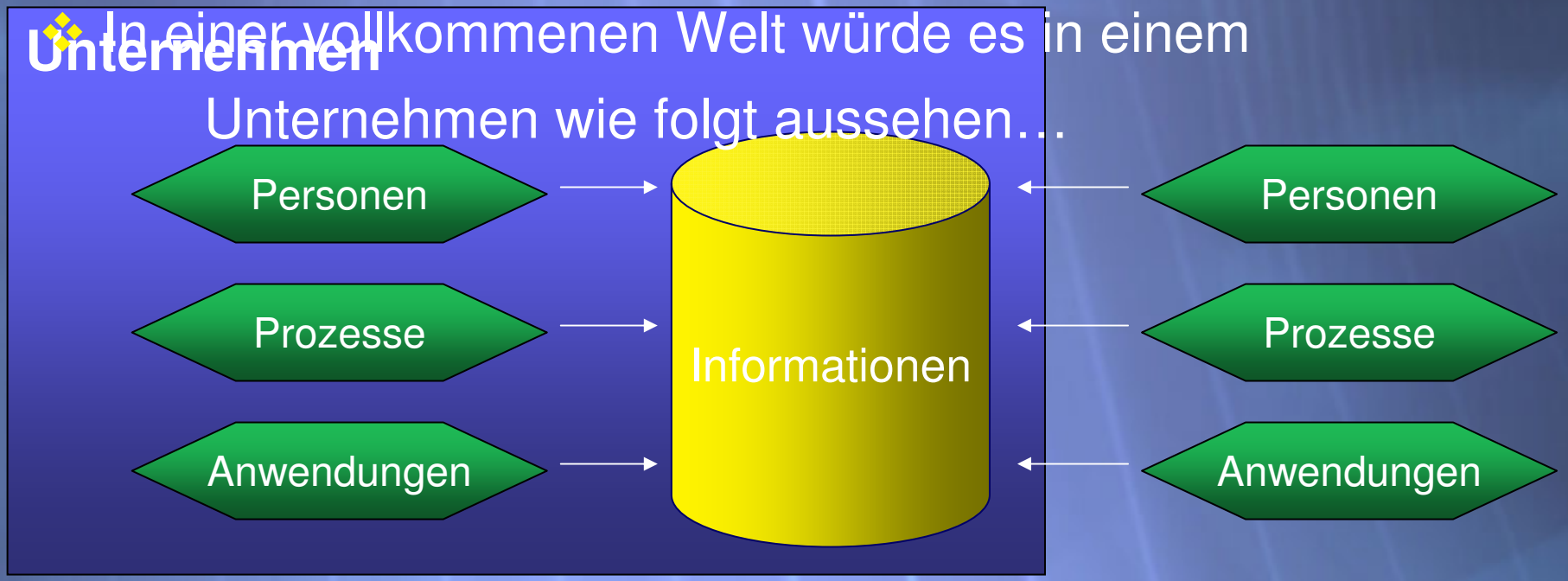
5. Problemstellung bei der Integration
6. Formen von Heterogenität
7. Integrationsmethoden und Prozesse
8. Implementierung in der realen Welt
9. Ausblick

1. Motivation

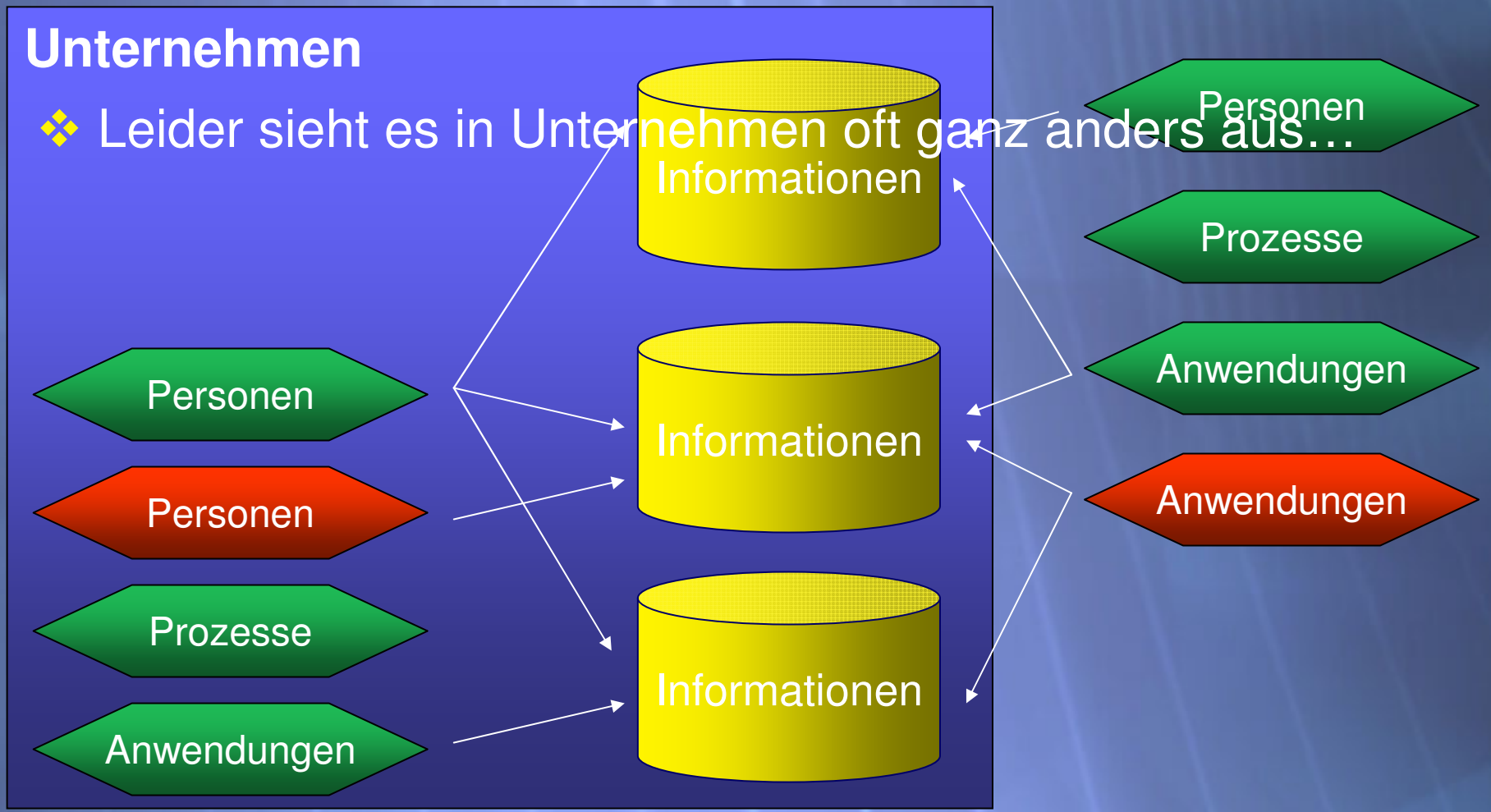
DATEN SIND ALLGEGENWÄRTIG …

Bill Inmon, "father of data warehousing"

- In einer vollkommenen Welt würde es in einem Unternehmen wie folgt aussehen…



1. Motivation



2. Die Informationsintegration als Teil der Unternehmensintegration

- **Problem:** Unabhängig funktionierende Informationssysteme im Unternehmen
- **Weg:** Vereinigung der Informationssysteme
- **Gründe:**
  - Enormes Einsparpotential
  - Steigerung der Arbeitsproduktivität
  - Rationalisierung der Unternehmensprozesse
  - Steigerung der Zusammenarbeit zwischen verschiedenen Abteilungen des Unternehmens

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Frage:** Was bedeutet Unternehmensintegration?
  (business integration)

- Autonome Systeme, Komponenten und Datenquellen im Unternehmen müssen weiter genutzt werden können und das Problem der Verteiltheit und Heterogenität dieser Daten muss gelöst werden.

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Herangehensweisen und Techniken bei der Unternehmensintegration - vier Kategorien**
  - Erstellung von Portalen

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Herangehensweisen und Techniken bei der Unternehmensintegration - vier Kategorien**
  - Erstellung von Portalen
  - Integration von Geschäftsprozessen

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Herangehensweisen und Techniken bei der Unternehmensintegration - vier Kategorien**
  - Erstellung von Portalen
  - Integration von Geschäftsprozessen
  - Integration von Anwendungen

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Herangehensweisen und Techniken bei der Unternehmensintegration - vier Kategorien**
  - Erstellung von Portalen
  - Integration von Geschäftsprozessen
  - Integration von Anwendungen
  - Informationsintegration (Datenintegration)

2. Die Informationsintegration als Teil der Unternehmensintegration

- **Unternehmensintegration** ist:
  - eine Kombination der vier vorgestellten Techniken
- **Informationsintegration** ist:
  - die Basis der Unternehmensintegration

3. Theorie der Datenintegration

Definitionen

- Theorie der Datenintegration $\subseteq$ Datenbanktheorie
- **Formal:**

$$< G , S , M >$$

  - Globales Schema (global)
  - Heterogene Menge der Quellenschemata (sources)
  - Menge der Abbildungen (mapping)

3. Theorie der Datenintegration

Definitionen

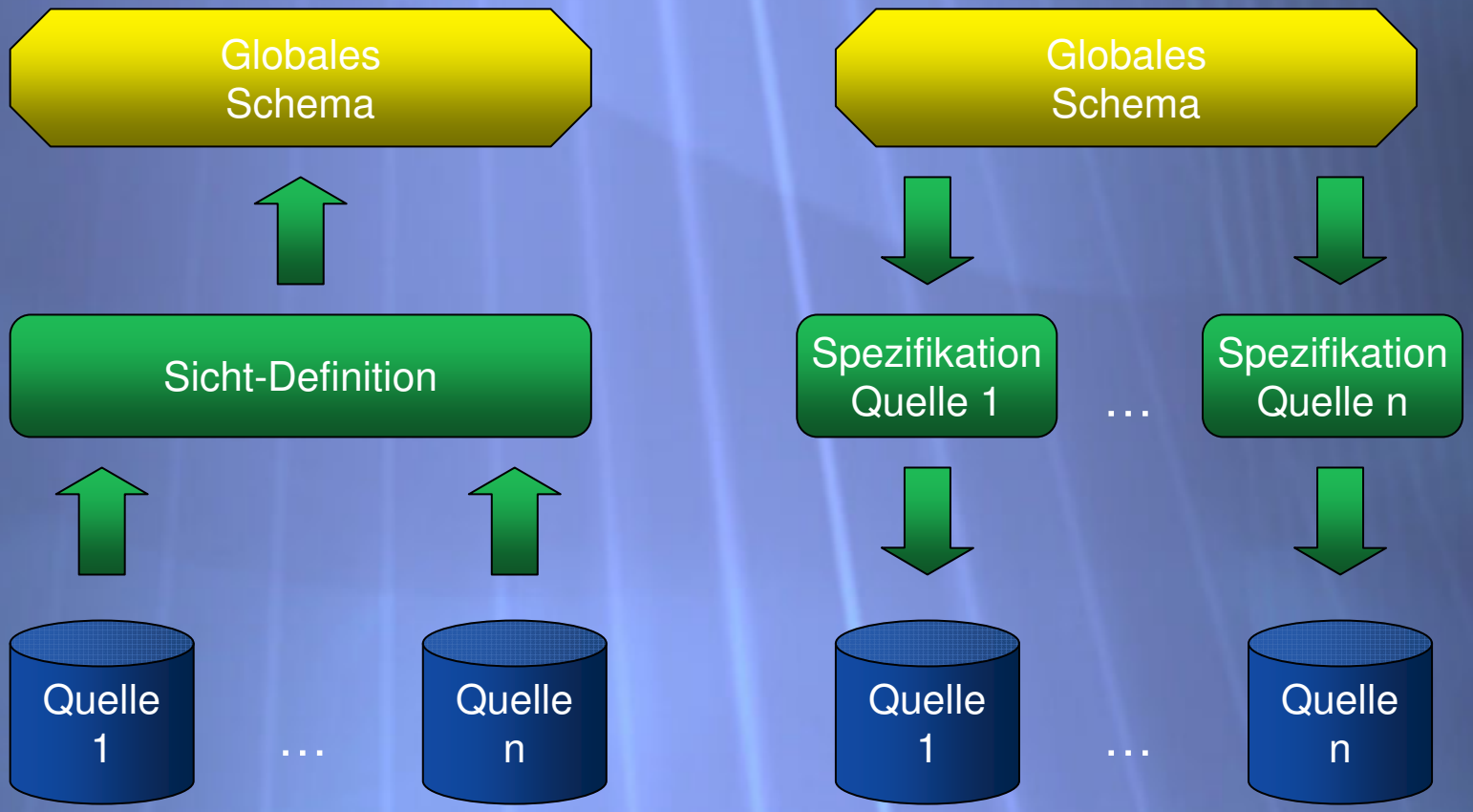
- Theorie der Datenintegration $\subseteq$ Datenbanktheorie
- **Formal:**

$$< G , S , M >$$

  - Globales Schema (global)
  - Heterogene Menge der Quellenschemata (sources)
  - Menge der Abbildungen (mapping)

3. Theorie der Datenintegration

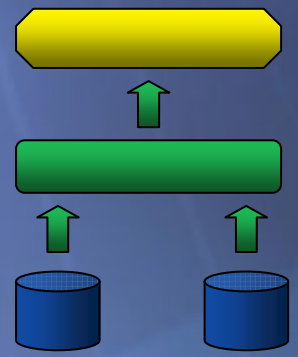
Global-as-View (GaV) vs. Local-as-View (LaV)



3. Theorie der Datenintegration

Global-as-View (GaV)

- M verbindet mit jedem Element aus G eine Anfrage auf S
- Abbildung zwischen G und S ist wohldefiniert
- **Nachteil:**
  - Komplizierte Implementierung des Vermittlercodes
  - Erweiterung sehr schwierig
- **Beispiel:**
  - Onlineshop



3. Theorie der Datenintegration
Global-as-View (GaV)

- **Beispiel:**

Gegeben ist ein globales Schema mit zwei Relationen:

- `Student: Matrikelnummer, Name, Alter, Semester`
- `Adresse: Matrikelnummer, Ort`

Diese sollen als sicht auf die folgenden lokalen Schemata dargestellt werden:

- `Q1: Matrikelnummer, Name, Ort`
- `Q2: Name, Matrikelnummer, Alter`
- `Q3: Matrikelnummer, Alter, Semester`

3. Theorie der Datenintegration

Global-as-View (GaV)

Für die Adressen kann nur Quelle `Q1` herangezogen werden

```
- CREATE VIEW Adresse AS
      SELECT Matrikelnummer, Ort
      FROM Q1
```

Studentendaten sind bis auf das Semester in der Quelle `Q2` und in Kombination in `Q1` und `Q3` enthalten:

```
- CREATE VIEW Student AS
      SELECT Matrikelnummer, Name, Alter FROM Q2
      UNION
      SELECT Q1.Matrikelnummer, Q1.Name, Q3.Alter
      FROM Q1, Q3
      WHERE Q1.Matrikelnummer = Q3.Matrikelnummer
```

3. Theorie der Datenintegration

Global-as-View (GaV)

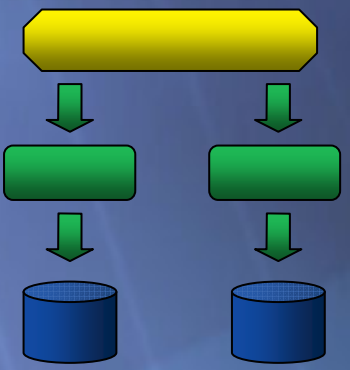
Falls beispielsweise nur Studenten aus Kaiserslautern berücksichtigt werden sollen, so ist die Sicht der Adressen:

```
- CREATE VIEW Adresse AS
      SELECT Matrikelnummer, Ort
      FROM Q1
      WHERE Ort="Kaiserslautern"
```

3. Theorie der Datenintegration

Local-as-View (LaV)

- Quellendatenbank als Menge von Sichten von G
- M bildet jede Relation von S auf eine Anfrage über G ab.
- **Nachteil:**
  - Beziehung ist nicht bijektiv → Abbildung von G auf S nicht wohldefiniert
  - Anfrageprozess kann lange dauern
- **Beispiel:**
  - Onlineshop



3. Theorie der Datenintegration

Local-as-View (LaV)

- **Beispiel:**

Gegeben sind drei lokale Datenquellen mit folgendem Schema:

- `Q1: Matrikelnummer, Name, Ort`
- `Q2: Name, Matrikelnummer, Alter`
- `Q3: Matrikelnummer, Alter, Semester`

Diese sollen auf folgendes globale Schema abgebildet werden:

- `Student: Matrikelnummer, Name, Alter`

3. Theorie der Datenintegration

Local-as-View (LaV)

Die Sichten der Quellen auf das globale Schema sind:

```
- CREATE VIEW S1 AS
      SELECT Matrikelnummer, Name
      FROM Student
- CREATE VIEW S2 AS
      SELECT Name, Matrikelnummer, Alter
      FROM Student
- CREATE VIEW S3 AS
      SELECT Matrikelnummer, Alter
      FROM Student
```

3. Theorie der Datenintegration

Local-as-View (LaV)

Sei im globalen Schema eine weitere Relation enthalten, die Matrikelnummern und Wohnorte einander zuordnet:

- `Adresse: Matrikelnummer, Ort`

Dann lässt sich die Quelle `Q1` darstellen als:

```
- CREATE VIEW S1 AS
      SELECT Student.Matrikelnummer, Student.Name,
             Adresse.Ort
      FROM Student, Adresse
      WHERE Student.Matrikelnummer =
             Adresse.Matrikelnummer
```

3. Theorie der Datenintegration

Local-as-View (LaV)

Die Quelle `Q2` enthält nur Studenten ab 25 Jahren, demnach entspricht die Sicht folgender:

```
- CREATE VIEW S2 AS
      SELECT Matrikelnummer, Name, Alter
      FROM Student
      WHERE Alter >= 25
```

3. Theorie der Datenintegration

Weitere Ansätze

- Global-Local-as-View
- Both-as-View

4. Mechanismen der Integration

Replikationsorientierte vs. Virtuelle Integration

- **Replikationsorientierte Integration**
  - Daten aus jeder Datenquelle in ein zentrales Datensystem kopieren und auf der zentralen Datenbank Anfragen ausführen
- **Virtuelle Integration**
  - Daten in ihren ursprünglichen Datenquellen belassen und bei Anfragen diese Datenquellen ansteuern

4. Mechanismen der Integration

Replikationsorientierte Integration

- **Exkurs:** Data Warehousing
  - Ein Data Warehouse ist eine Datenbank, welche
    - ✓ themenorientiert ist

Bill Inmon, “father of data warehousing”

4. Mechanismen der Integration

Replikationsorientierte Integration

- **Exkurs:** Data Warehousing
  - Ein Data Warehouse ist eine Datenbank, welche
    - ✓ themenorientiert ist
    - ✓ zeitvariant ist
    - ✓ permanent ist
    - ✓ integriert ist

Bill Inmon, “father of data warehousing”

4. Mechanismen der Integration

Replikationsorientierte Integration

- **Durchführung:** Extract, Transform, Load Prinzip (ETL)
  - Extrahieren der Daten aus ihren ursprünglichen Datenquellen
  - Ablegen der Daten in einem einheitlichen Format
  - Hochladen der Daten in das Data Warehouse

4. Mechanismen der Integration

Replikationsorientierte Integration

- **Vor- und Nachteile:**
  - \+ Integrationschritte nur einmal durchzuführen
  - \+ Schnelle Ausführung der Anfragen
  - − Speicheraufwändig
  - − Schwieriges Updaten der Daten
  - − Lokale Anwendungen arbeiten oft auf nicht aktuellen Daten

4. Mechanismen der Integration

Virtuelle Integration

- Daten bleiben in Ihren ursprünglichen Datenquellen
- Zentrale Portale bilden den Zugang zu den einzelnen Datenquellen
- Anfragen werden umgewandelt, so dass die verschiedenen verteilten Informationssystem sie verarbeiten können

4. Mechanismen der Integration

Virtuelle Integration

- **Vor- und Nachteile:**
  - \+ Speichereffizient
  - \+ Die lokalen Daten bleiben aktuell
  - − Laufzeitineffizient

5. Problemstellung bei der Integration

- **Ziele der Integration:**
  - Middlewaresysteme (MWS) bereitstellen, welche
    - ✓ Daten wie lokale Daten betrachten
    - ✓ anspruchsvolle Such-, Transformations- und Analysedienste anbieten
    - ✓ Interaktion mit anderen MWS unterstützen

5. Problemstellung bei der Integration

- **Herausforderungen bei der Datenintegration:**
  - Datenheterogenität

5. Problemstellung bei der Integration

- **Herausforderungen bei der Datenintegration:**
  - Datenheterogenität
  - Vereinigung und Verteilung von Daten

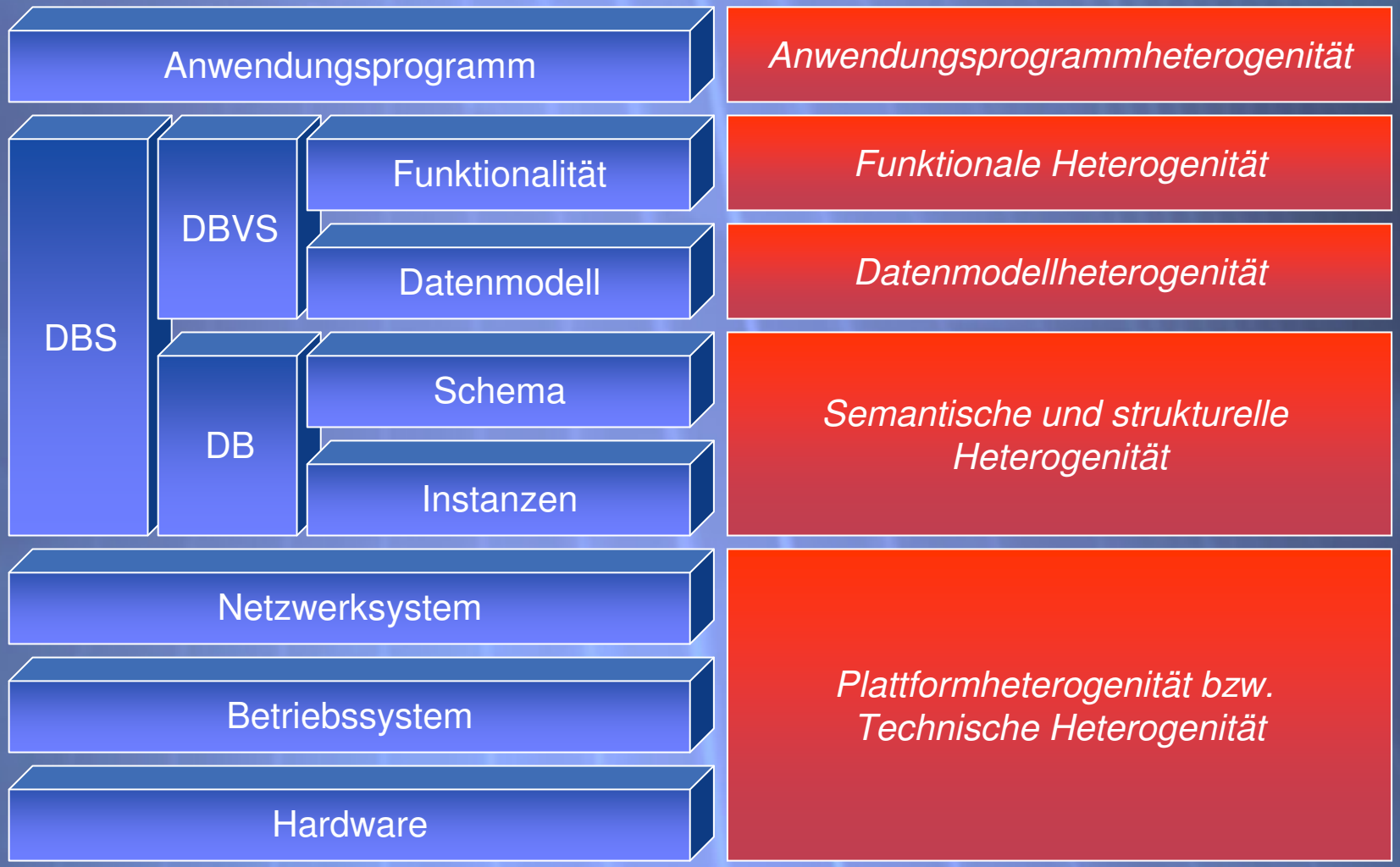
5. Problemstellung bei der Integration

- **Herausforderungen bei der Datenintegration:**
  - Datenheterogenität
  - Vereinigung und Verteilung von Daten
  - Schaffen von *business intelligence*

  → Hohe Ansprüche an das **Informationsintegrations-verwaltungssystem**

  → **XML** und **Web Services** von immer größerer Bedeutung

6. Formen der Heterogenität



6. Formen der Heterogenität

6.1. Datenmodellheterogenität

- Daten gemäß verschiedener Modelle abgespeichert
- **Konzepte:** Aggregationen, Generalisierung, Assoziationen, …
- **Beispiel:** Generalisierung im objekt-orientierten Datenmodell, nicht vorhanden im relationalen Datenmodell
  - Daten strukturell sehr unterschiedlich abgelegt obwohl sie denselben Sachverhalt darstellen

6. Formen der Heterogenität

6.2. Semantische Heterogenität

- Daten werden nach logischem Schema abgelegt
- **Problem:** was ist logisch?
- **Beispiel:** Schema zum Abspeichern von Namensdaten:

  <[NAME], [VORNAME]>

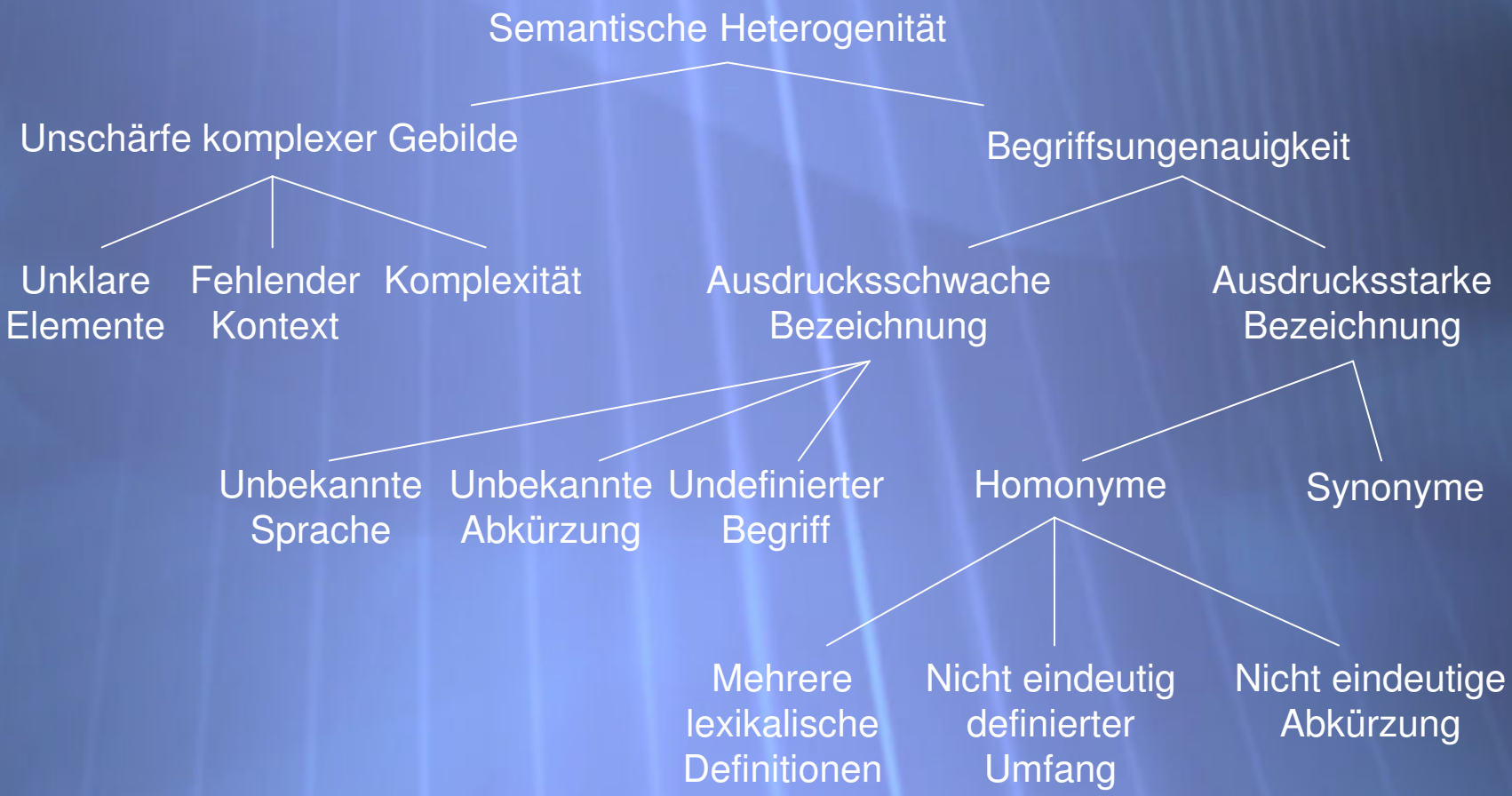
  <[VORNAME], [NAME]>

  Wie ist folgender Eintrag zu interpretieren?

  <Hans,Thomas>

6. Formen der Heterogenität

6.2. Semantische Heterogenität



6. Formen der Heterogenität

6.2. Semantische Heterogenität

- **Probleme:**
  - Fehlende Dokumentation erschwert den Integrationsvorgang
  - Unterschiedliche Definitionsbereiche
  - Integrationsvorgang nicht voll-automatisierbar

6. Formen der Heterogenität

6.3. Strukturelle Heterogenität

- **Beispiel:** Personen werden mit Vor-, Nachname und Geschlecht in einer Tabelle abgelegt
  - Mögliche Darstellungen:

Tabelle A: männliche Personen

Vorname	**Nachname**
Peter	Schmidt

Tabelle B: weibliche Personen

Vorname	**Nachname**
Johanna	Wunder

6. Formen der Heterogenität

6.3. Strukturelle Heterogenität

- **Beispiel:** Personen werden mit Vor-, Nachname und Geschlecht in einer Tabelle abgelegt
  - Mögliche Darstellungen:

Tabelle A: Personen

Vorname	Nachname	männlich	weiblich
Peter	Schmidt	1	0
Johanna	Wunder	0	1

6. Formen der Heterogenität

6.3. Strukturelle Heterogenität

- **Beispiel:** Personen werden mit Vor-, Nachname und Geschlecht in einer Tabelle abgelegt
  - Mögliche Darstellungen:

Tabelle A: Personen

Vorname	Nachname	Geschlecht
Peter	Schmidt	männlich
Johanna	Wunder	weiblich

6. Formen der Heterogenität

6.3. Strukturelle Heterogenität

- **Beispiel:** Ablegen eines Buches in einer Datenbank

Tabelle A: Buch (Schema 1)

ID	Titel	Autor_ID	Autor
12345	Datenbankanwendungen	1	Härder

Tabelle A: Buch (Schema 2)

ID	Titel	Autor_ID
12345	Datenbankanwendungen	1

Fremdschlüsselbeziehung →

Tabelle B: Autor (Schema 2)

Autor_ID	Autor
1	Härder

- ✓ Gleiche Elemente eine gemeinsamen Datenmodells
- ✓ Unterschiedliche Darstellung

7. Integrationsmethoden und -prozesse

- Integration in 2 Phasen:
  - **Phase I:**
    - Interpretieren von Daten
    - Bereinigen von Daten
    - Transformieren von Daten
    - Verbinden von Daten
  - **Phase II:**
    - Verbinden von verschiedenen Datensystemen

7. Integrationsmethoden und -prozesse

7.1. Interpretieren von Daten

- Durchkämmen und Interpretieren der zu integrierenden Daten
- **Beispiel:** <Peter Schmidt, 37 Jahre alt, ledig>
  <Schmidt Peter, 37>
  - **Interpretation:** Es handelt sich zweimal um die gleiche Person!

7. Integrationsmethoden und -prozesse

7.2. Bereinigen von Daten

- Identifikation, Standardisierung, Anpassung und Befreiung von Redundanzen
- **Beispiel:** <Peter Schmidt, 37 Jahre alt, ledig>
  <Schmidt Peter, 37>
  - **Bereinigung:** Identifikation und Löschen des redundanten Eintrags <Schmidt Peter, 37>

7. Integrationsmethoden und -prozesse

7.3. Transformieren von Daten

- Umwandlung der Daten in benötigtes Format und Qualität
- **Beispiel:** `<Peter Schmidt, 37 Jahre alt, ledig>`
  - **Transformation:** Speichern des Dateneintrages als

    `<Schmidt, Peter, 37, ledig>`

7. Integrationsmethoden und -prozesse

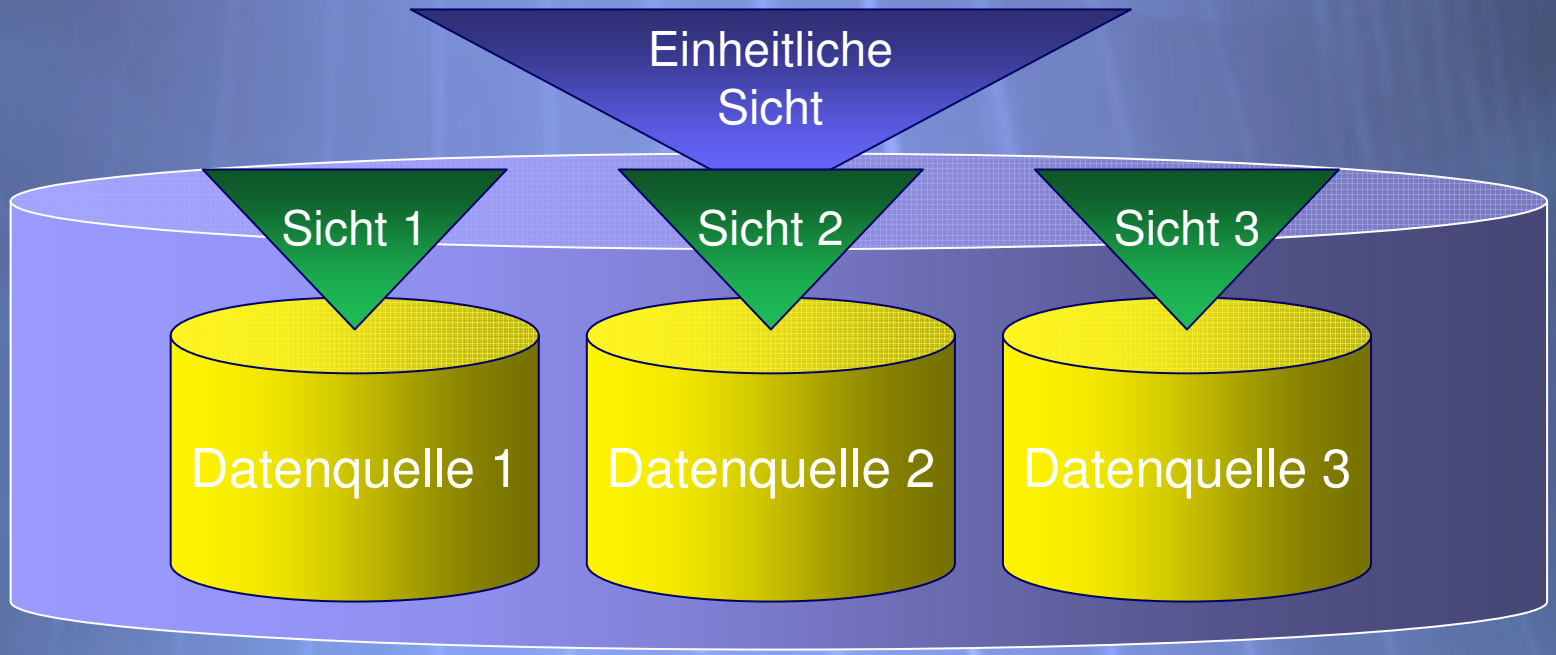
7.4. Verbinden der Daten

- Verbinden von zusammen gehörenden Daten
- **Beispiel:** `<Schmidt, Peter, 37, ledig>`
  `<Raabengasse, 67657, Kaiserslautern>`
  - **Verbinden:** Peter Schmidt, 37 Jahre alt, ledig wohnt in der Raabengasse, 67657 Kaiserslautern.

7. Integrationsmethoden und -prozesse

7.5. Verbinden von verschiedenen Datensystemen

- Verbinden von mehreren Datenquellen



8. Implementierung in der realen Welt

Integration bei der Wachovia Corporation

- Mittlerweile die 4. größte Bank in den USA
- 2001: Fusion mit der First Union Bank
- 2002: Übernahme der PRB
- 2004: Fusion mit der South Trust Corp.
- Von 21.000 auf 100.000 MA in 6 Jahren
- Vielzahl isolierter Informationssysteme

➢ **Probleme für:** IT Abteilung, Servicepersonal

8. Implementierung in der realen Welt

Integration bei der Wachovia Corporation mit WebSphere

- Virtuelle Integration
- Zentrales Anfrageportal
- Erstellung von Desktopanwendungen
- Mühelose Interaktion der Mitarbeiter mit "dem Informationssystem"
- Lokale Anwendungen können immer noch auf aktuellen Daten arbeiten

➢ **Einsparungen in Millionenhöhe**

8. Implementierung in der realen Welt

Integration bei der Wachovia Corporation mit WebSphere

9. Ausblick

- Die verschiedenen Konzepte sind noch immer nicht ganz ausgereift und werden ständig weiterentwickelt
- Datenintegration wird durch wirtschaftliche Lage und Globalisierung von immer größerer Bedeutung

… DATEN SIND ALLGEGENWÄRTIG

Bill Inmon, “father of data warehousing”

Fragen…?

…Antworten!