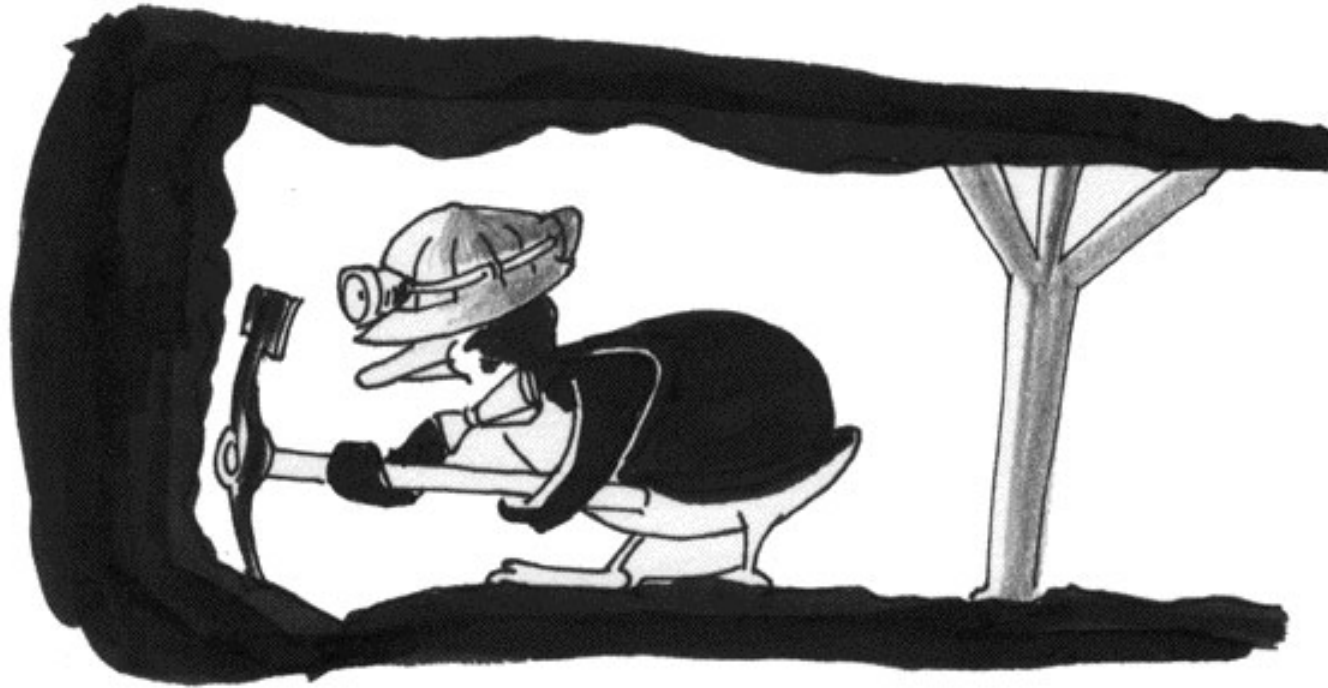


Seminar Business Intelligence Teil II

Data Mining & Knowledge Discovery



Was ist Data Mining?

Sabine Queckbörner

- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP
- Problemfelder des Data Mining

- **Begriffsbestimmung**
 - Analyse großer Datenmengen mit automatischen oder halbautomatischen Werkzeugen (Berry, Linoff)
 - Methode zur Problemlösung um logische oder mathematische, zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken (Decker, Focardy)

- **Begriffsbestimmung**

- Teilschritt des KDD-Prozesses, der aus der Anwendung von Datenanalysealgorithmen besteht und zu einer Auflistung von Mustern, die aus Daten gewonnen wurden, führt (Fayyad, Piatetsky-Shapiro und Smyth)

⇒ Data Mining als Anwendung von Algorithmen auf Daten mit der Zielsetzung, Muster aus den Daten zu extrahieren.

Data Mining

Nach welchen Mustern wird gesucht

- **Data Mining**
 - Was ist Data Mining?
 - **Nach welchen Mustern wird gesucht?**
 - Wie wird gesucht?
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP

Data Mining

Nach welchen Mustern wird gesucht

- Ausprägungen
 - Regeln und Abhängigkeiten
 - Gruppen (Cluster)
 - Verbindungsmuster (Link)
 - Zeitliche Muster (Sequence)
 - Abweichungen
 - Formeln und Gesetzmäßigkeiten

Data Mining

Nach welchen Mustern wird gesucht

- ✘ Nicht alle gefundenen Muster sind für den Benutzer interessant!
- ✘ Nicht alle für den Benutzer relevanten Muster werden gefunden!

- **Data Mining**
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - **Wie wird gesucht?**
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP
- Problemfelder des Data Mining

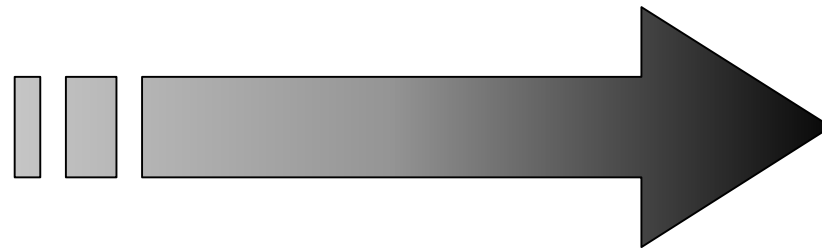
- Möglichkeiten
 - Nach mehreren Mustern parallel suchen
 - In verschiedenen Abstraktionsebenen suchen
 - Nach verschiedenen Schwerpunkten

- Methoden
 - Künstliche neuronale Netze
 - Kohonen-Netze
 - Lineare Regression
 - Genetische Algorithmen
 - CHAID
 - Regelbasierte Systeme

- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- **Knowledge Discovery in Databases**
 - **Der KDD-Prozess**
 - Gegenüberstellung Data Mining und KDD
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP
- Problemfelder des Data Mining



Daten



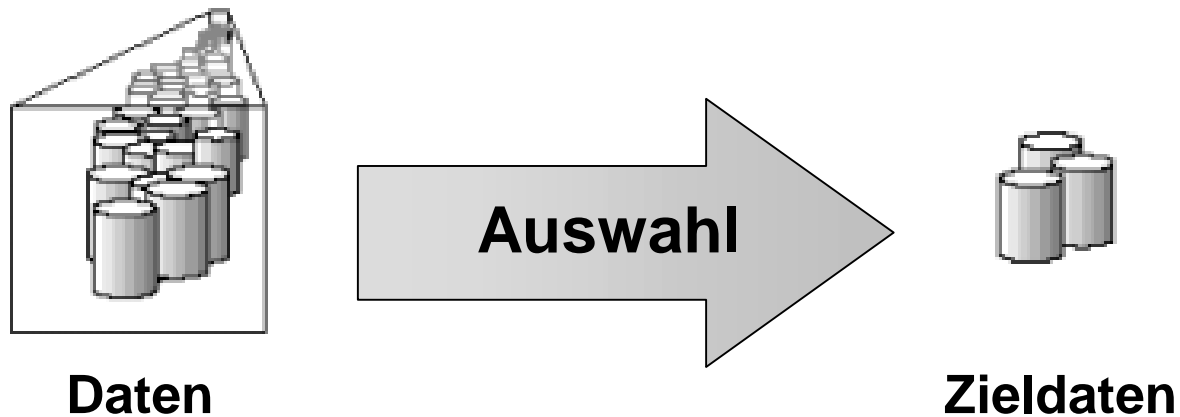
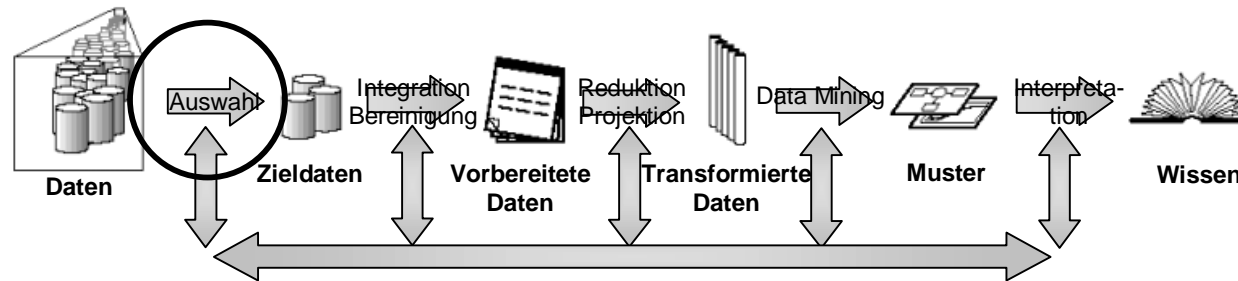
Wissen

Prozess zur Identifikation von Mustern

- Gültig
- Neuartig
- Potentiell nützlich
- Verständlich

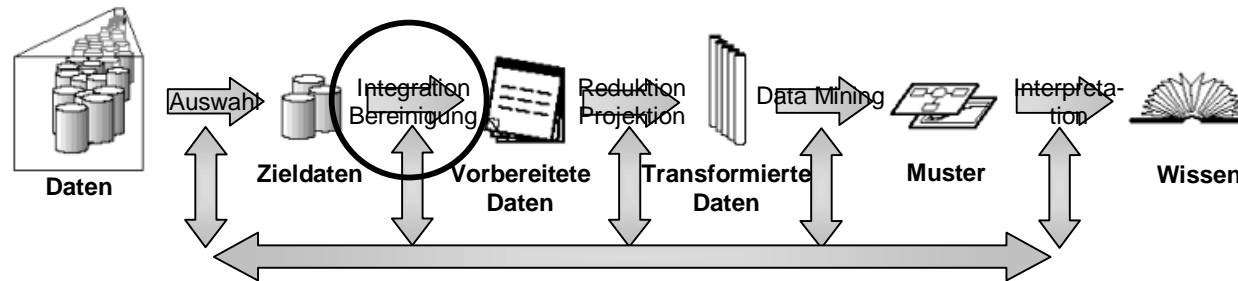
KDD

Der KDD-Prozess



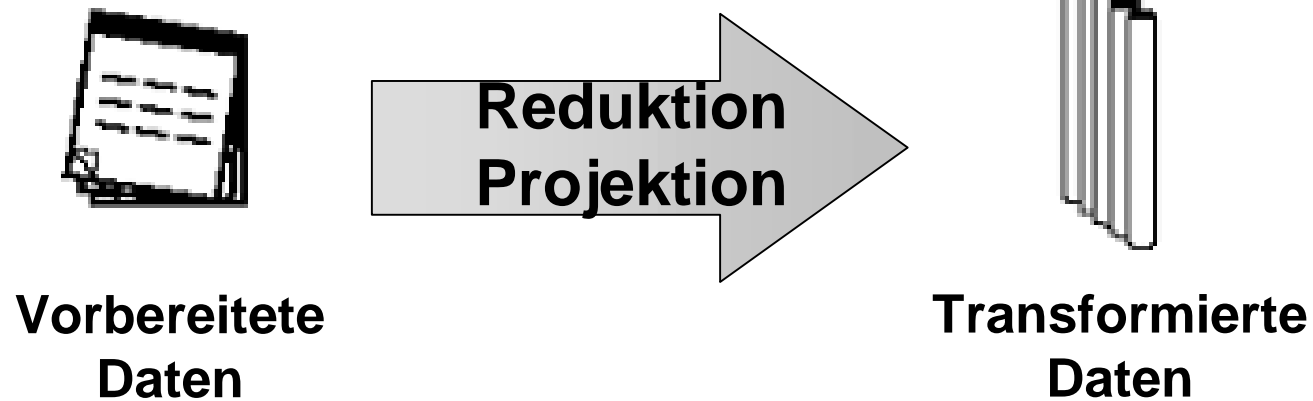
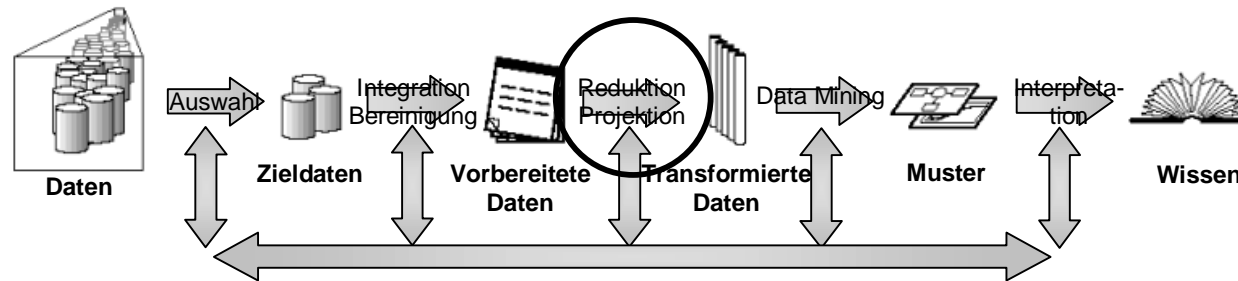
KDD

Der KDD-Prozess



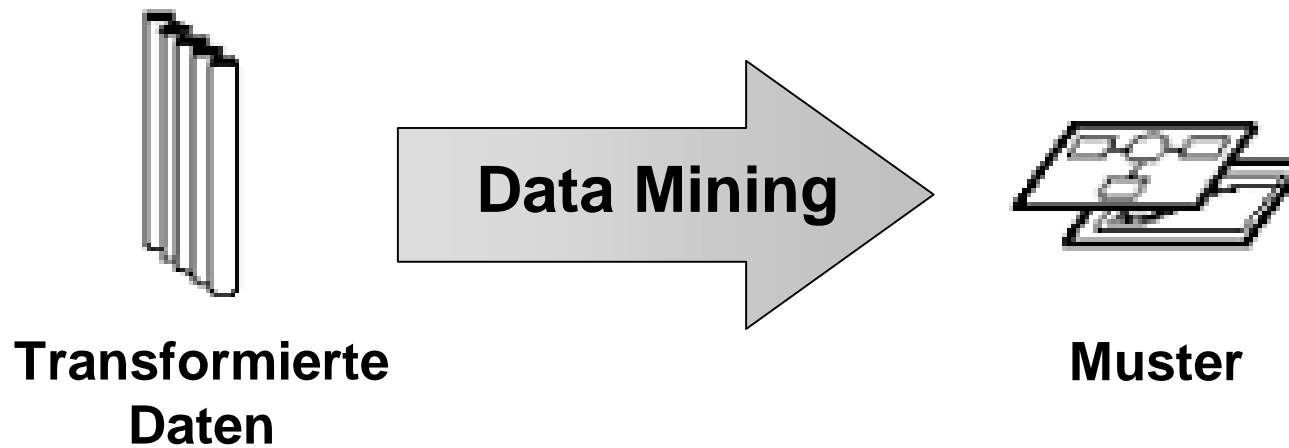
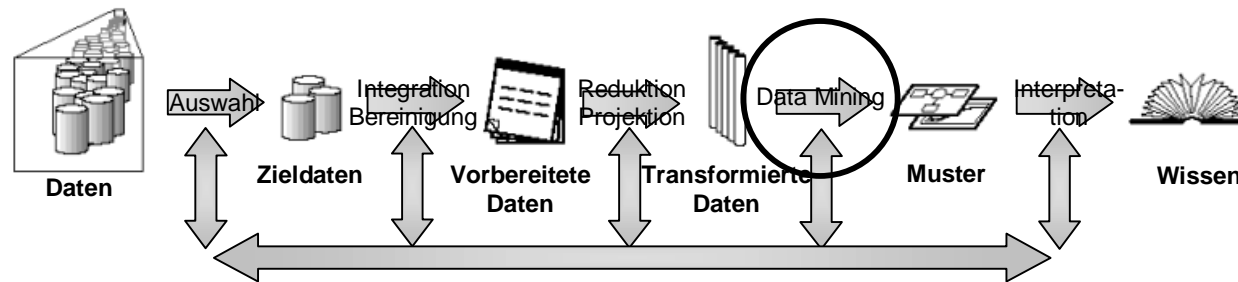
KDD

Der KDD-Prozess



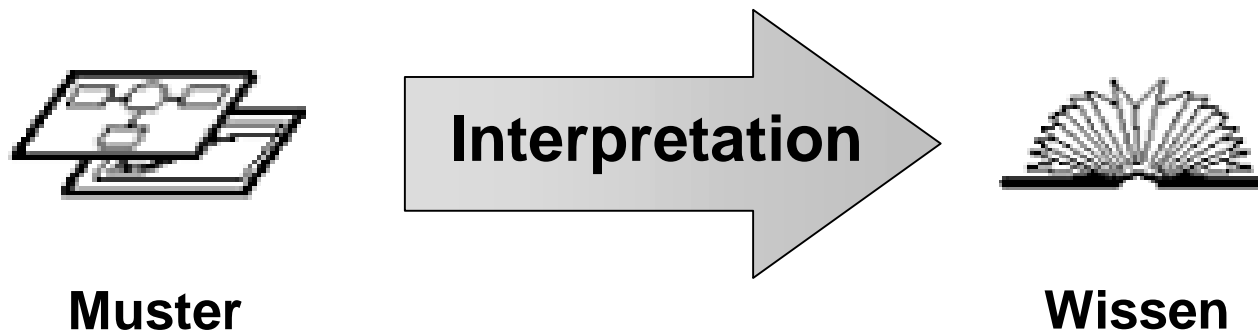
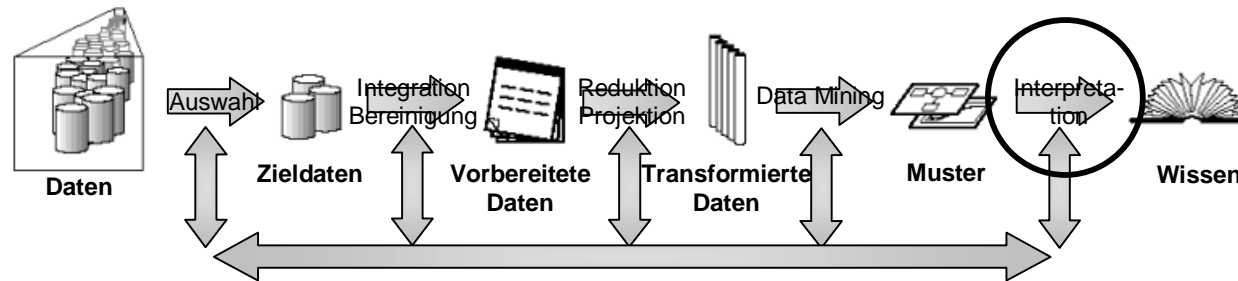
KDD

Der KDD-Prozess



KDD

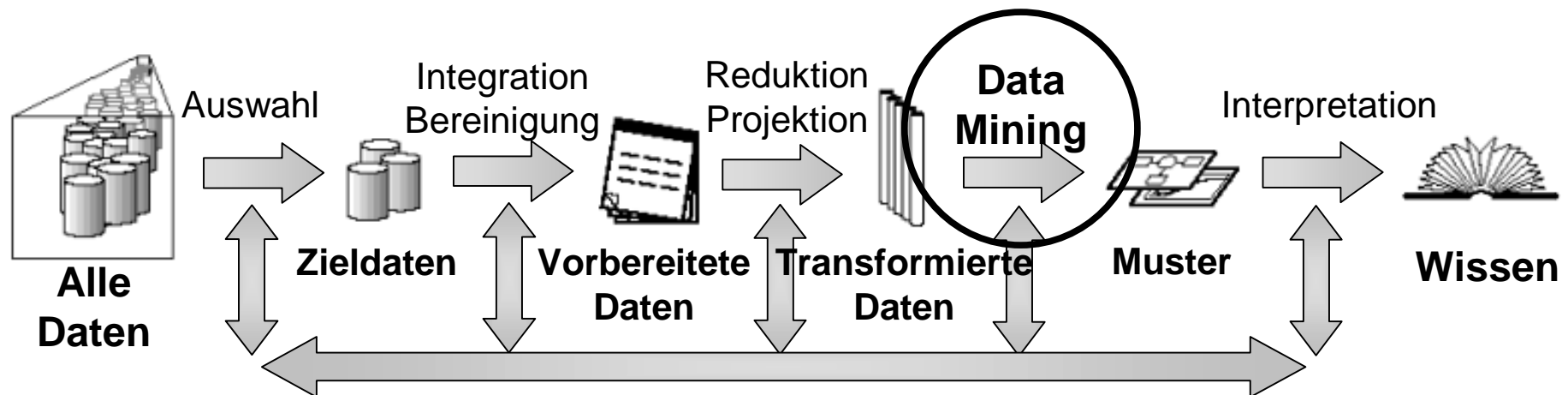
Der KDD-Prozess



- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- **Knowledge Discovery in Databases**
 - Der KDD-Prozess
 - **Gegenüberstellung Data Mining und KDD**
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP
- Problemfelder des Data Mining

KDD

Data Mining \Leftrightarrow KDD



✘ Data Mining ist ein Teilschritt des KDD-Prozesses

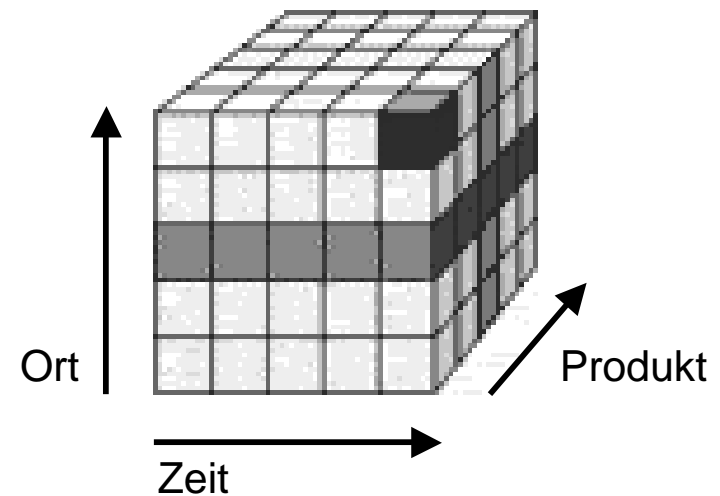
- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- **Online Analytical Processing (OLAP)**
 - **OLAP-Funktionen**
 - Gegenüberstellung Data Mining und OLAP
- Problemfelder des Data Mining

- Datenwürfel
 - n Dimensionen
 - ⇔ n Attribute

Beispiel:

3 Dimensionen:

⇒ Ort, Zeit, Produkt



- Datenwürfel

- n Dimensionen

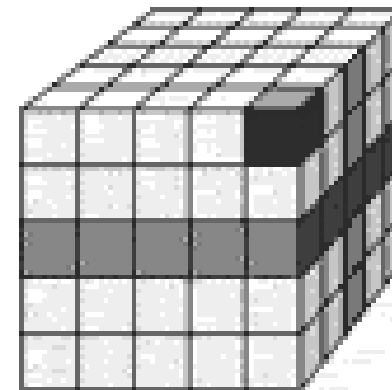
⇔ n Attribute

- Wertebereich

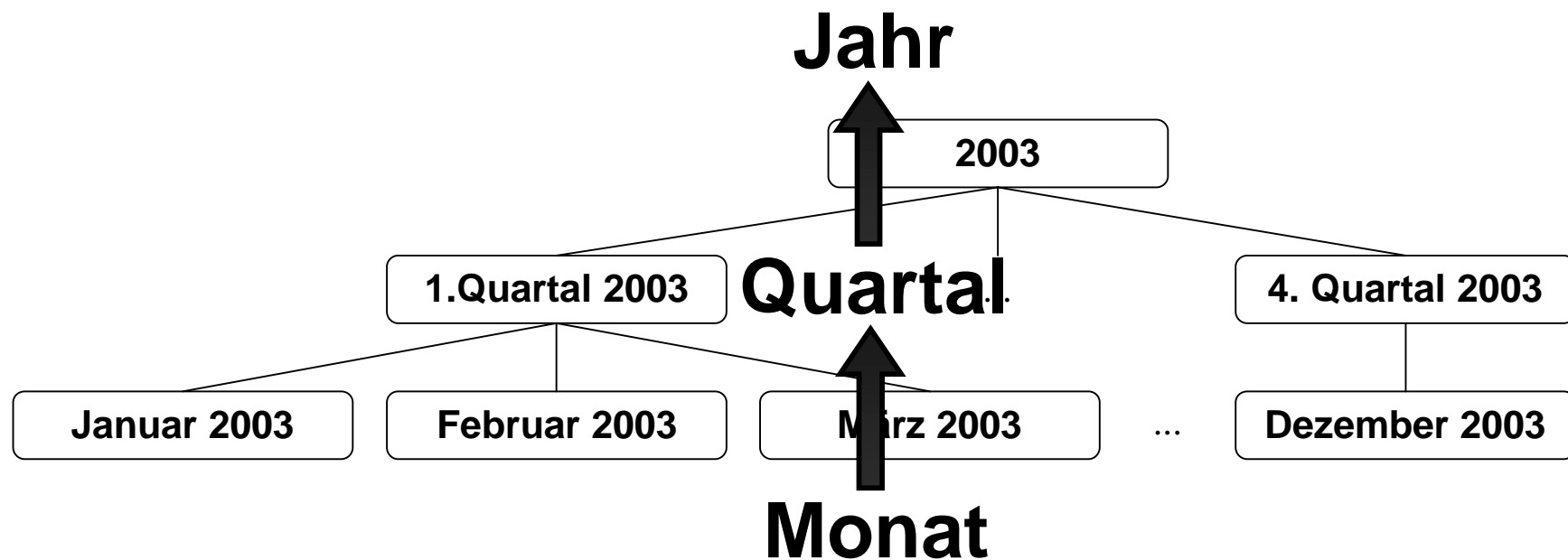
kontinuierlich ⇔ diskret

- Hierarchien

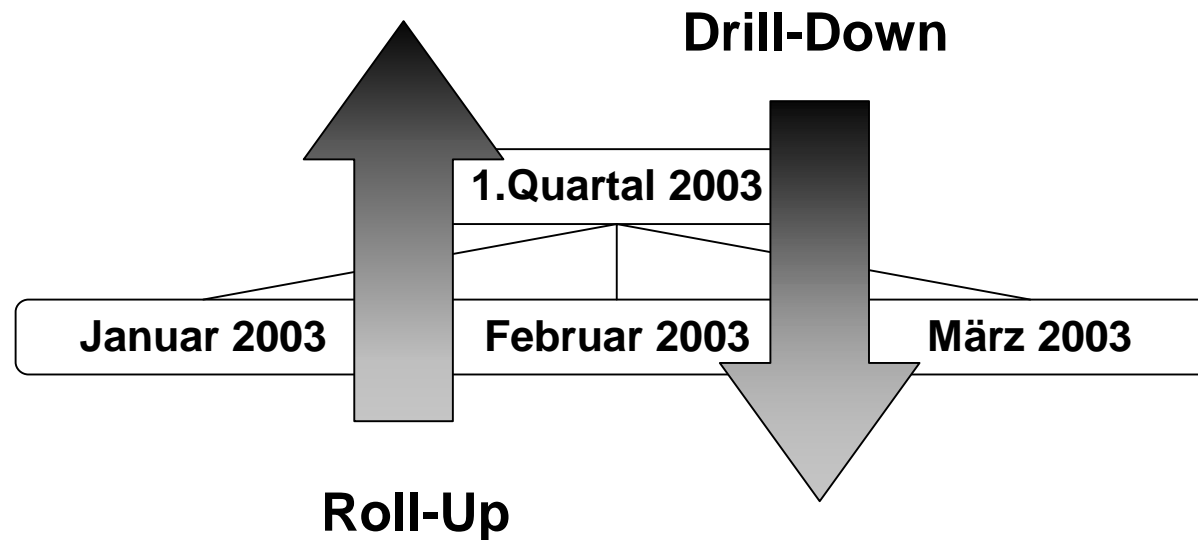
⇒ Abstraktionsebenen



- Hierarchien
 - Beispiel: *Hierarchie auf Attribut „Zeit“*



- OLAP-Funktionen
 - Roll-Up
 - Drill-Down

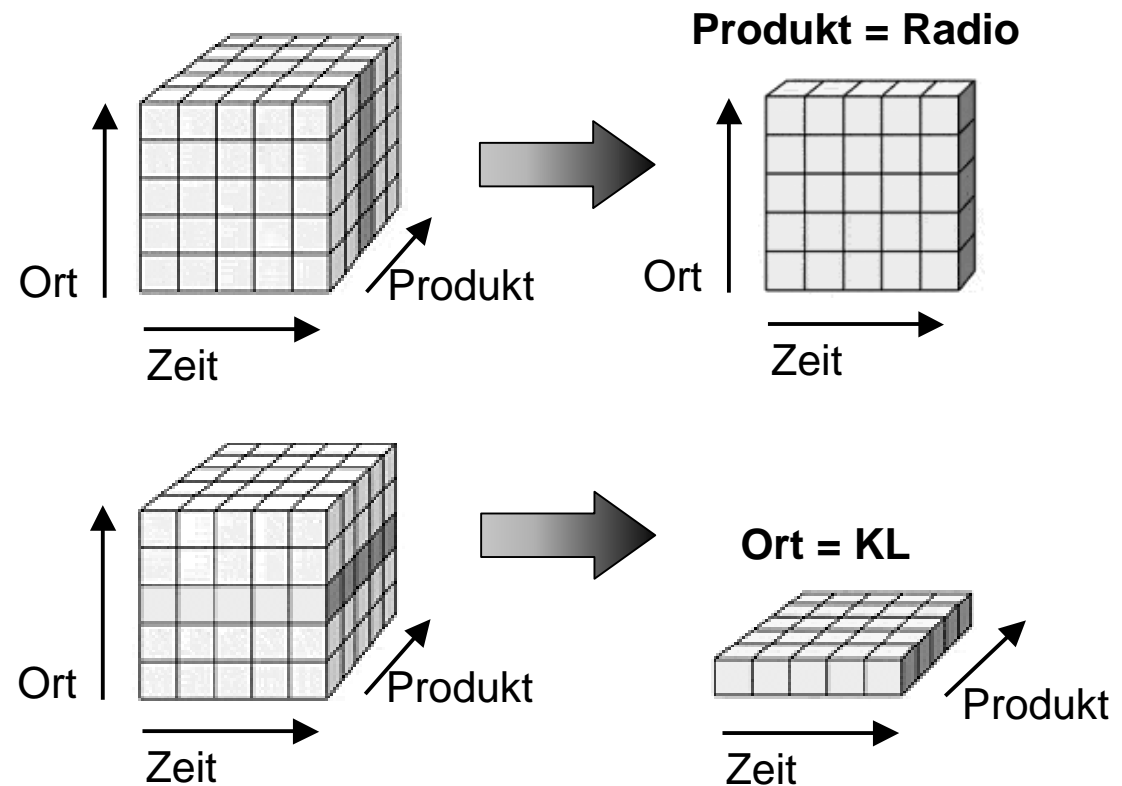


OLAP

OLAP-Funktionen

● OLAP-Funktionen

- Roll-Up
- Drill-Down
- Slice



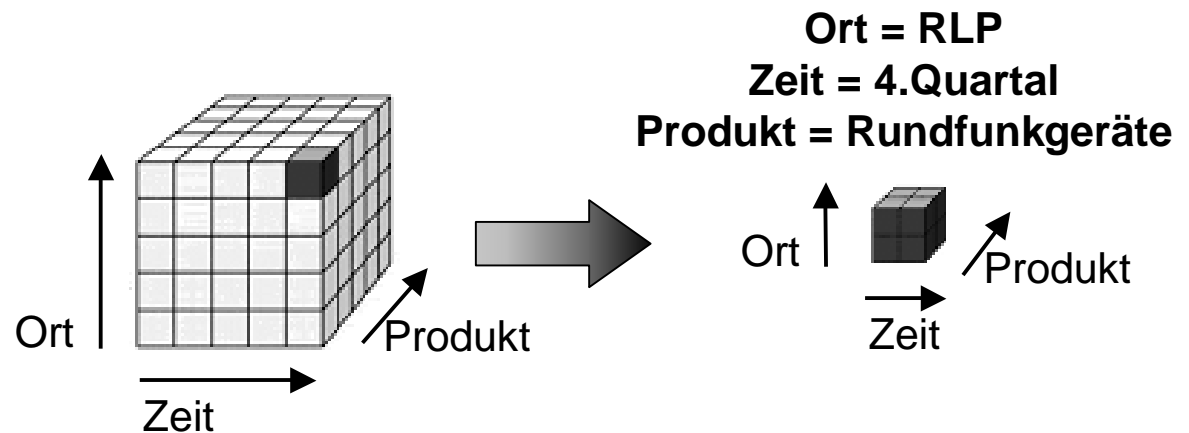
OLAP

OLAP-Funktionen

- OLAP-Funktionen

- Roll-Up
- Drill-Down

- Slice
- Dice



- OLAP-Funktionen
 - Roll-Up
 - Drill-Down

 - Slice
 - Dice
 - Rotate (Pivote)

OLAP

Data Mining \Leftrightarrow OLAP

- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- **Online Analytical Processing (OLAP)**
 - OLAP-Funktionen
 - **Gegenüberstellung Data Mining und OLAP**
- Problemfelder des Data Mining

OLAP

Data Mining \Leftrightarrow OLAP

- Rennpferd \Leftrightarrow Arbeitspferd
 - OLAP = Arbeitspferd
 - Data Mining = Rennpferd

OLAP

Data Mining \Rightarrow OLAP

● Data Mining als Erweiterung des OLAP

OLAP	\Rightarrow	Data Mining
x Analyseziel notwendig (z.B. Hypothese)	\Rightarrow	Kein Analyseziel notwendig (kann Hypothesen liefern)
x Ergebnis steht fest (z.B. Bestätigung der Hypothese)	\Rightarrow	Ergebnis steht nicht fest, unbekannte Muster möglich
x Daten müssen bekannt sein (Benutzer wählt aus)	\Rightarrow	Daten können vollkommen unbekannt sein
x Benutzerinteraktive Analyse (aktive Auswertung der Daten)	\Rightarrow	Automatische Analyse (computergestützte Algorithmen)
x Ergebnis ist eine andere Darstellungsform der Daten	\Rightarrow	Ergebnis sind Zusammenhänge und Muster in den Daten

OLAP

Data Mining \Leftarrow OLAP

- OLAP Mining
 - integrierte, konsistente und gesäuberte Daten
 - Anwendung von Data-Mining-Algorithmen
 - Interaktive Analyse und anschauliche Interpretation wie bei OLAP
- OLAP-Mining-Funktionen (Auszug)
 - Cubing then Mining
 - Mining then Cubing

Problemfelder

Problemfelder des Data Mining

- Data Mining
 - Was ist Data Mining?
 - Nach welchen Mustern wird gesucht?
 - Wie wird gesucht?
- Knowledge Discovery in Databases (KDD)
 - Der KDD-Prozess
 - Gegenüberstellung Data Mining und KDD
- Online Analytical Processing (OLAP)
 - OLAP-Funktionen
 - Gegenüberstellung Data Mining und OLAP
- **Problemfelder des Data Mining**

Problemfelder

Problemfelder des Data Mining

● Speichergröße

- Problem: Daten passen nicht vollständig in den Speicher
- Lösung: Algorithmen, die zum Beispiel in mehreren Durchgängen analysieren

● Antwortzeit

- Problem: schnellere Analysen gewünscht – kürzere Antwortzeiten des Systems
- Lösung: parallele Analysetechniken

● Vielfalt der Datentypen

- Problem: neue Datentypen, Videos ...

● Benutzerfreundlichkeit

● Datenschutz und Sicherheit

Was ist Data Mining?

Zusammenfassung

- Data Mining
 - Muster, Verfahren
- KDD-Prozess
 - Data Mining als Teilschritt des KDD-Prozesses
- OLAP
 - Datenmodell, OLAP-Funktionen
 - Data Mining als Erweiterung des OLAP
- Problemfelder

Vielen Dank