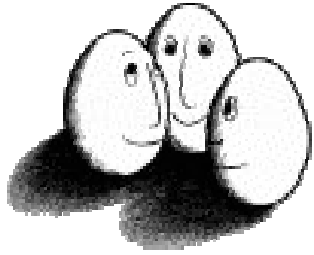


# **PG-402 Wissensmanagement: Zeitreihen in Datenbanken**

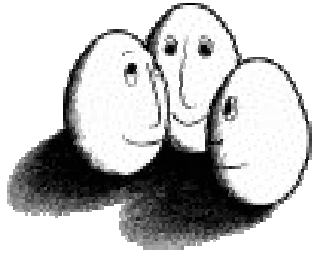
07.05.2002

Klaus Unterstein



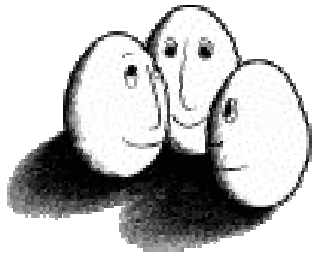
# Übersicht

- Zeitreihen
- Schritte bei der Zeitreihenanalyse
- Anwendungsfallbeispiel: SLTIME2
- Zusammenfassung
- Anhang: Literaturangaben



# I. Zeitreihen

- Was sind Zeitreihen ?
- Wer braucht sie ?
- Wo findet man sie ?
- Warum gerade Zeitreihen ?
- Wofür sind sie gut ?

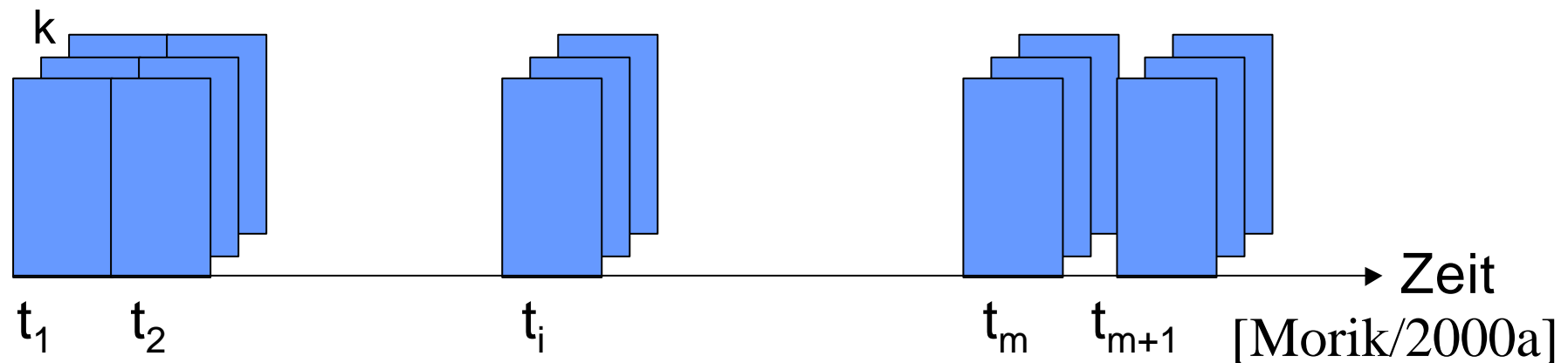


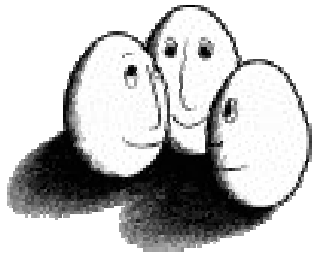
# I. Zeitreihen

**Univariat** - ein Attribut pro Zeit (Temperatur)

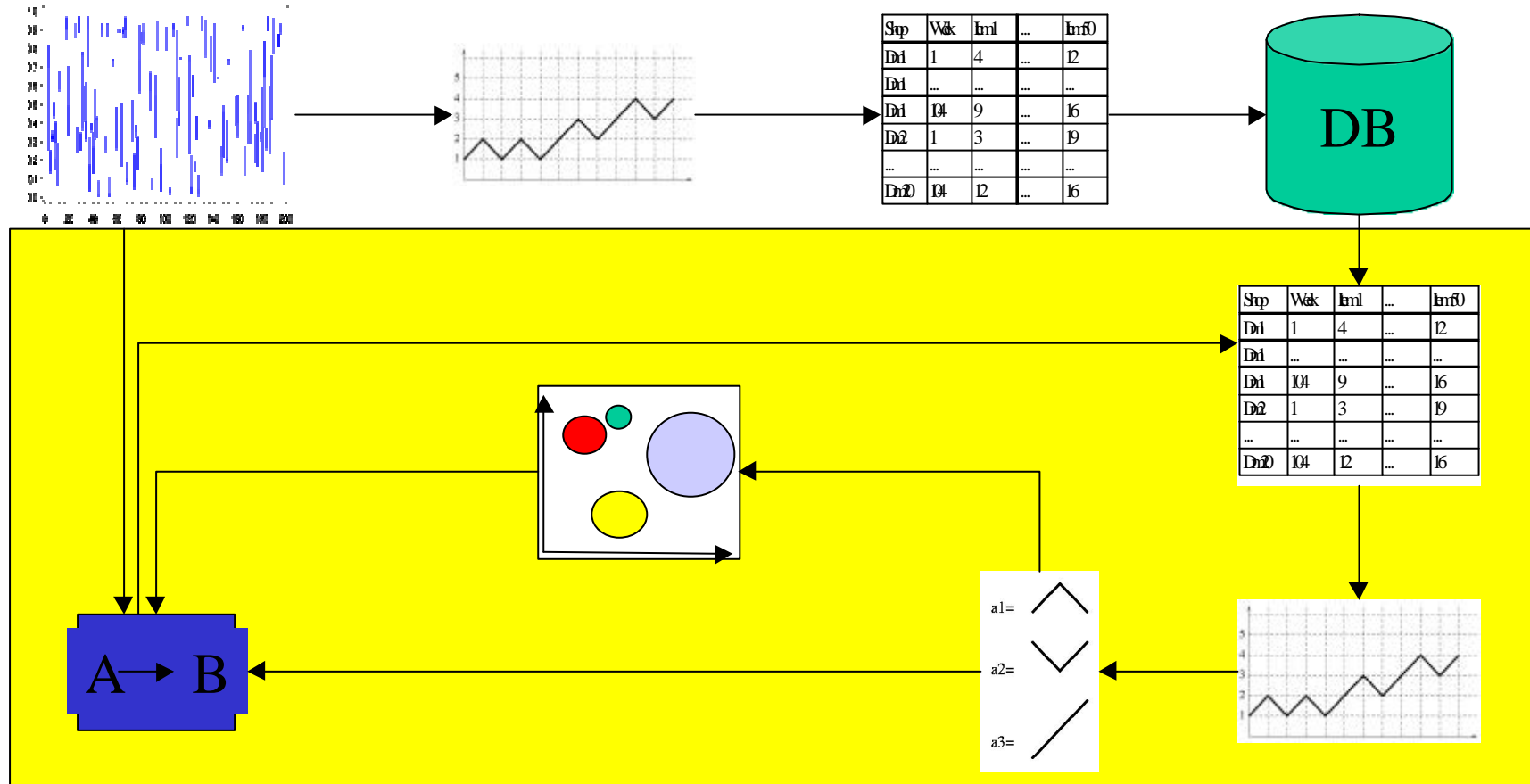


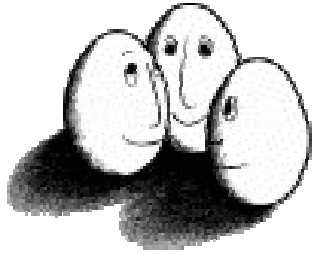
**Multivariat** - k Attribute (Temperatur, Luftdruck, Luftfeuchtigkeit)





# II. Schritte bei der Zeitreihenanalyse

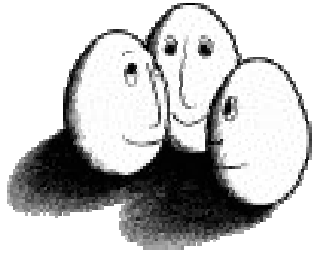




## II. Zeitreihenanalyse

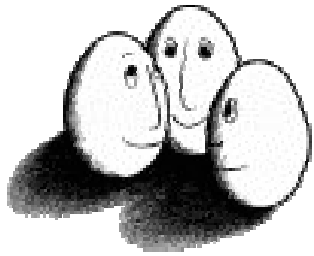
### Reihenfolge

- Messung
- `Tabellarisierung`
- Speicherung
- Daten-Vorverarbeitung
- Diskretisierung
- Clustering
- Assoziationsregeln
- Bewertung
- Iterationen



## II. Zeitreihe – Schritte

- Messung
- `Tabellarisierung` (bereits eine Daten-  
Vorverarbeitung zur Speicherung)  
Beispiel: Data-Warehouse
- Speicherung

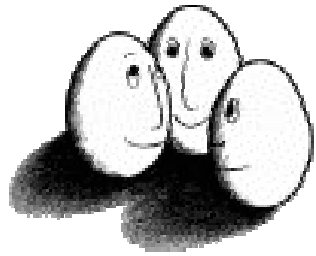


## II. Zeitreihe – Schritte (2)

- Daten-Vorverarbeitung für Zeitreihenanalyse (Data-Preprocessing)
  - Eignung der Daten für eine bestimmte Aufgabe (Qualitäts- und Quantitätsprüfung)
  - Festlegung der Lösungsmenge
  - Wahl des Verfahrens
  - Aneignung von Domain-Wissen
  - Dateninspektion (Behandlung v. Problemen, NULL-Einträge,...)
  - Vereinfachung(en) für das gewählte Verfahren

[nach: DP/1999 mit Modifikationen]

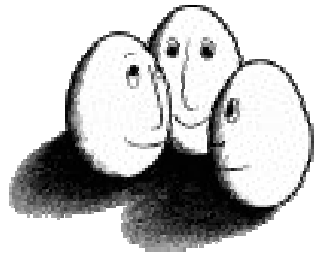




## II. Zeitreihe – Schritte (3)

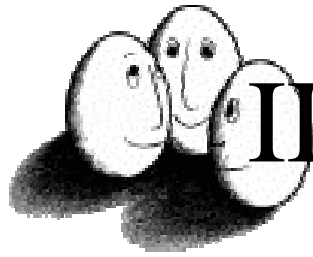
- Diskretisierung von Zeitreihen
  - Umwandlung von kontinuierlichen Größen in diskrete Größen
  - Gegeben: Sequenz ( $s$ )  
 $s = (x_1, \dots, x_n)$
  - Wahl der Fenstergröße ( $w$ )
  - Subsequenz der Zeitreihe ( $s_i$ )  
 $s_i = (x_i, \dots, x_{i+w-1})$

[Nach: DLM/1998]



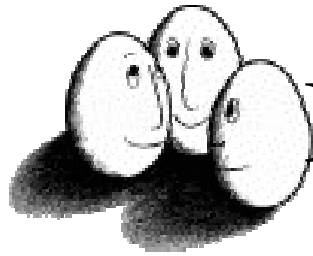
## II. Zeitreihe – Schritte (4)

- (Clustering der gefundenen Sub-Sequenzen [k-Means, Greedy-Algorithmen,...])
- Nach Assoziationsregeln suchen auf der `diskreten` Zeitreihe
- Bewertung der erzeugten Regeln (manuell, J-Measure,...)
- Iterationen mit anderen Parametern, Verfahren, etc.



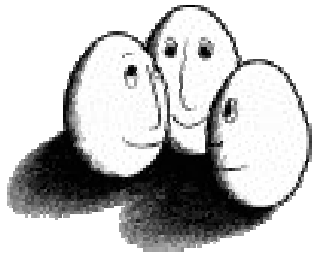
## III. Anwendungsfallbeispiel: SLTIME2

- Bisher: Ein (optimaler) Ablauf dargestellt
- Idealfall ohne Probleme – Praxis meist mit Komplikationen
- Probleme können in jedem Schritt auftreten



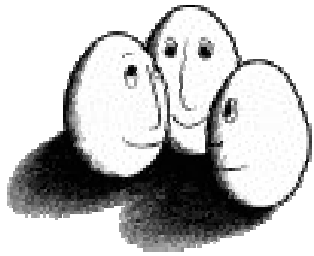
## III. Ergebnisse anhand der SLTIME2

- Schritt (1) – (3) waren unbeeinflussbar.
- (4) Datenvorverarbeitung:
  - Domainwissenaneignung (Verständnis der Strukturen)
  - `einfache` Bereinigung/Korrektur der Daten, Entdeckung von Problembereichen (NULLs, Inkonsistenzen, unmögliche Werte)
  - Transformation der Zeitstruktur für ein Verfahren
- (5) - (9) Scheitert derzeit an (4)



# III. Ergebnisse - Datentransformation

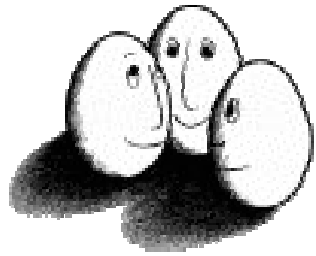
- Transformation: Alle Vertragsereignisse einer Person als Zeitreihe
  - Granularität der Daten
  - Vertragszuordnung (VVID->PTID)
  - Betrachtung der Änderungen
  - unterschiedliche Anzahl von zu betrachtenden Teilen
  - grosse Flexibilität des Systems
  - Kodierung der Ereignisse [  $f(x)$ -> Wert ]
  - NUR Label vs. Kompletter 'Spaltendump'
  - MANUELLE Umwandlung einzelner Person



## III. Ergebnisse – Datentransformation (2)

- Transformation:  
<Beginn, Label, Ende> in DB enthalten.  
Entscheidungen, wie unterschiedliche  
Versionen gehandhabt werden sollen.
- Weitere Ideen:  
- Shop-Beispiel

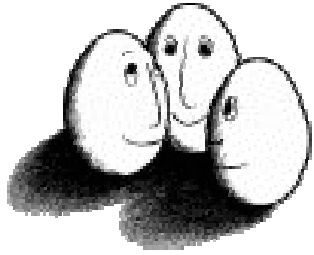
[Höpp/2001, HK/2001]



### III. Ergebnisse – Datentransformation (3)

Shop	Week	Item1	...	Item50
Dm1	1	4	...	12
Dm1	...	...	...	...
Dm1	104	9	...	16
Dm2	1	3	...	19
...	...	...	...	...
Dm20	104	12	...	16

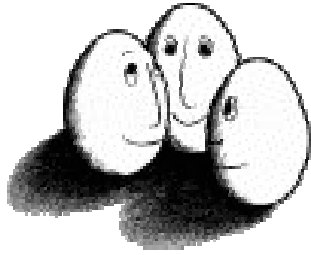
[Morik/2000a]



## IV. Zusammenfassung

- Gezeigt wurde ein allgemeiner Ablauf einer optimalen Zeitreihenanalyse
- Der aktuelle Stand der SLTIME2 wurde gewählt, um Beispiele und Probleme aufzuzeigen.
- Weiterhin wurden zwei Ansätze angesprochen, um in der nachfolgenden Diskussion die Machbarkeit und Alternativen zu überlegen!





# V. Anhang: Literaturangaben

- [DP/1999] Dorian Pyle. *Data Preparation for Data Mining*.  
Morgan Kaufmann, 1999
- [Morik/2000a] Katharina Morik. *The Representation Race – Preprocessing for Handling Time Phenomena*. In ECML‘2000
- [DLM/1998] Gautam Das, King-Ip Lin, Heikki Mannila. *Rule discovery from time series*
- [FB-2001] Fabian Bauschulte. *PG402 – Zeitaspekte*. 24.10.2001.
- [SG/2001] Ron Sun, C. Lee Giles. *Sequence Learning: From Recognition and Prediction to Sequential Decision Making* IEEE-2001
- [HK/2001] Frank Höppner, Frank Klawonn. *Finding Informative Rules in Interval Sequences*. In IDA‘2001.
- [Höpp/2001] *Learning Temporal Rules from State Sequences*. In WLTS‘2001.
- [GuDa] Dimitrios Gunopulos, Gautam Das. *Time Series Similarity Measures*.
- [MaSe] Heikki Mannila, Jouni K. Seppänen. *Finding similar situations in sequences of events via random projections*.

☺ Danke! ☺