KSFE

KONFERENZ DER SAS®-ANWENDER IN FORSCHUNG UND ENTWICKLUNG

KSFE 2003

BAUER

VERLAGSGRUPPE

Heuristische Verfahren zur Aggregation addierbarer Zeitreihen

BAUER

VERLAGVERTRIEBS KG

VKG Verlagvertriebs KG
Abt. Data Mining

Stefan Pohl
Dr. S. Steinberg

Agenda

- Heinrich Bauer Verlag
- Typisches Problem des Vertriebs
- Lösungsansatz: Supervised Clustering
- Vergleich verschiedener Verfahren
- SAS/EM-Knoten: „Cluster Comparison“
- Vergleich von Clusterscorings oder Klassifikationen
- Zusammenfassung

BAUER VERLAGSGRUPPE

Heinrich Bauer Verlag

- Europas führender Zeitschriftenverlag
- weltweit:
  - ca. 6300 Mitarbeiter
  - 120 auflagenstarke Erfolgstitel
- in Deutschland Marktführer in den Segmenten:
  - Programmzeitschriften (55% = 9,8 Mio. Exemplare)
  - Jugendzeitschriften (39% = 1,4 Mio. Exemplare)
  - unterhaltende Frauenzeitschriften (35% = 4,0 Mio. Exemplare)
- Special Interest Magazine

BAUER VERLAGSGRUPPE

VKG

Vertriebsfirma der Verlagsgruppe Bauer

- 730 Mio. € Umsatz
- 150 Mitarbeiter
- ca. 20 Mio. Hefte pro Erscheinungsintervall
- 3,8 Mio. Abonnenten
- jährlich ca. 800.000 neue Abonnenten
- HH Abdeckung durch VKG Logistik ca. 80%
- über 215 Mio. Liefertakte p.a. Logistik
- ca. 30% WBZ und Fremdtakte

Typisches Problem des Vertriebs

- Problem:

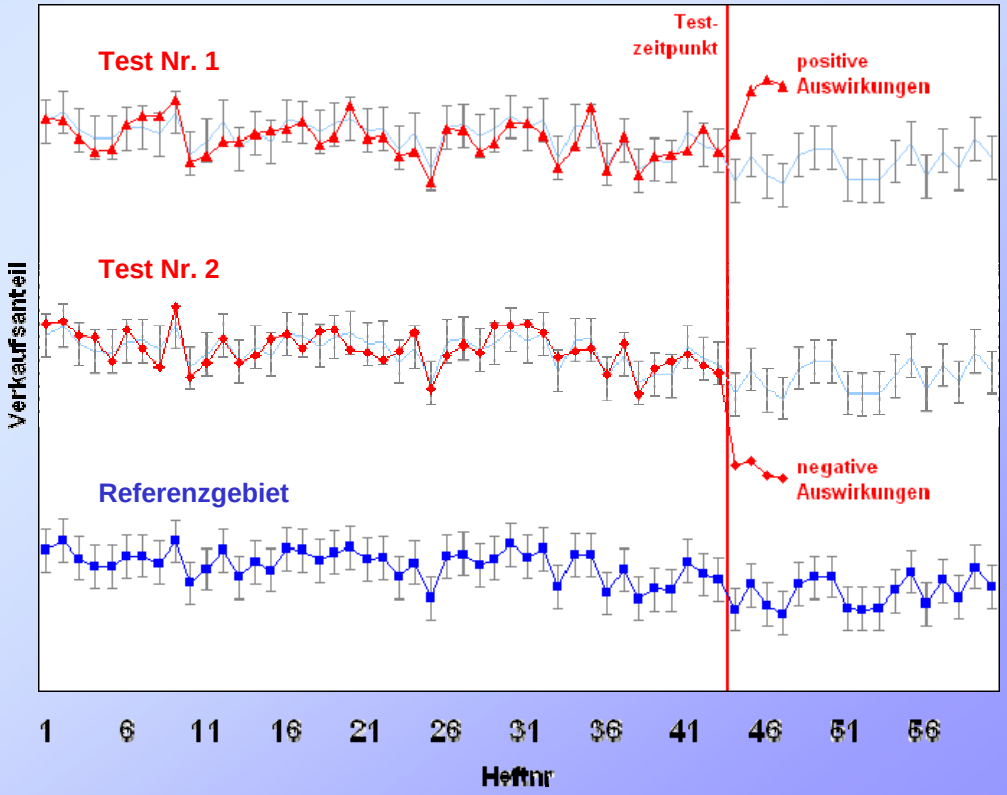
  Wie können Auswirkungen von vertrieblichen Aktivitäten
  - Bezugsänderungen
  - Werbemaßnahmen
  - …

  nachgewiesen werden?

- Lösungsansatz

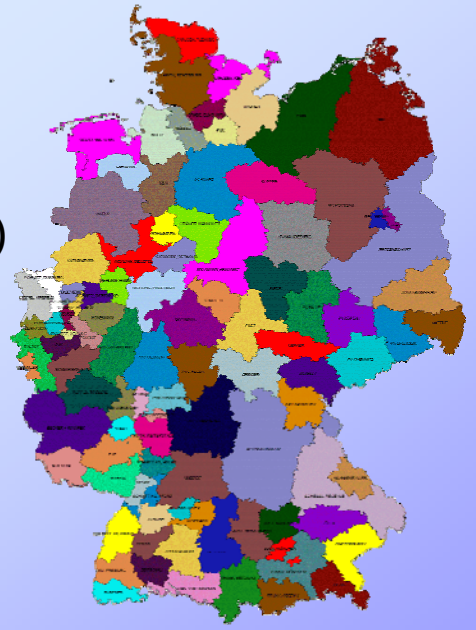
  Einteilung der Vertriebsgebiete in Test- & Referenz-gebiete

Zeitreihe von Verkaufsanteilen



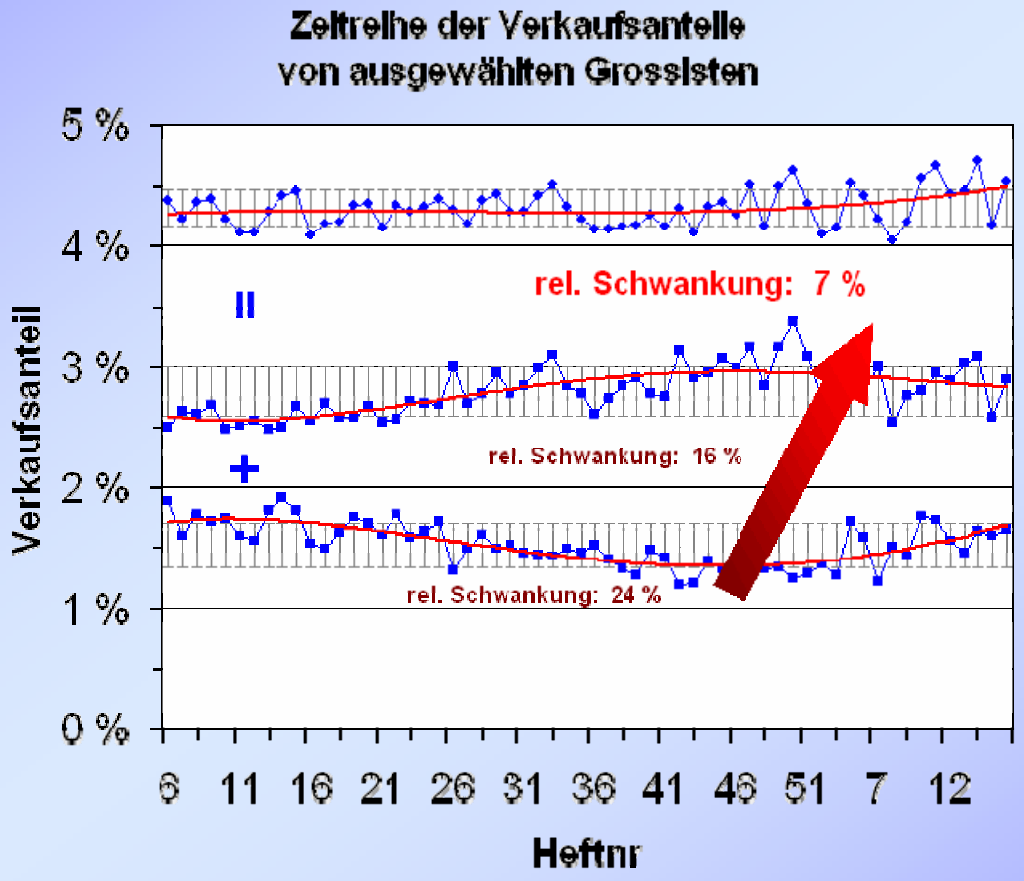
Konkretisierung

- Begriffe
  - Verhalten =
    Verkäufe als einzige Rückmeldung im Einzelhandel
  - Einflüsse =
    Inhalte der Zeitschrift,
    Titelbild (Hintergrundfarbe, Model, ...),
    Regionalität (Ost/West, Groß-/Kleinstadt, Wetter, ...)
  - Vertriebsgebiete =
    logistisch bedingte Einteilung Deutschlands
- Probleme
  - Titelthemen können regionale, zeitlich begrenzte Verkaufsschlager sein
  - Verkäufe abhängig von Affinitäten der Kunden,
    beliebig geographisch verteilt
  - Einfluß bisheriger Testaktivitäten verzerrt Historie



Lösung

- Zusammenfassung von Vertriebsgebieten zu Clustern, die ein besseres Verhalten haben



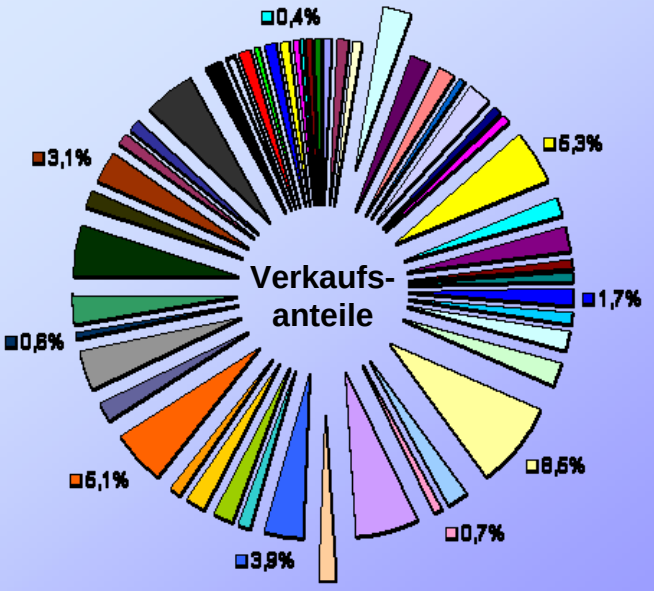
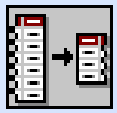


Abbildung auf Verfahren des SAS/EM

- Zufallsauswahl

  keine Optimalität erwartbar,
  kann aber gut als Referenz dienen

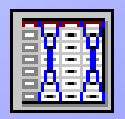


- Clustering

  Probleme:
  - Statik der Elementeigenschaften
  - Elemente erhalten neue Eigenschaften bei jedem Iterationsschritt
  - unabhängig von der aktuellen Clusterzusammenstellung
  - unüberwacht

- Anlehnung an Variablenselektion

  teilt Variablen in die beiden Gruppen
  „erklärend“ und „redundant“ ein



Anforderungen

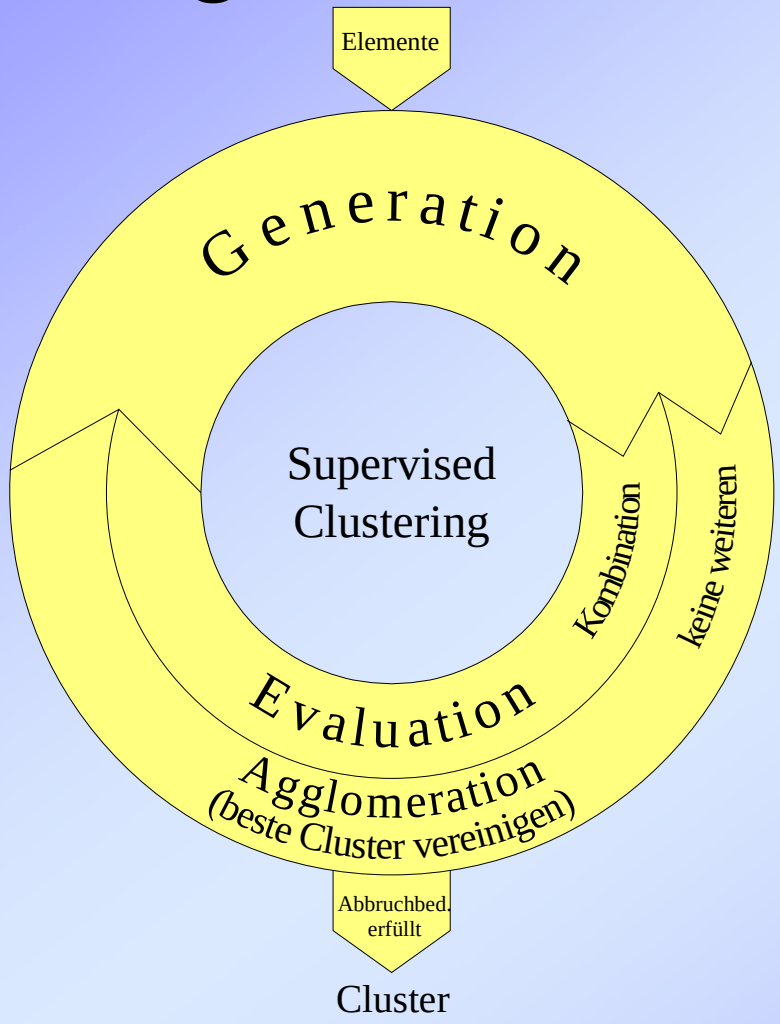
… an Testumgebung:

- mehrere Tests
- unterschiedliche Aktivitäten
  (Bezugsänderung, Werbung, ...)
- gleichzeitige Durchführbarkeit
- statistisch gesicherter Erkenntnisgewinn
- nicht nur: gab es Einfluß,
  sondern: wie hoch war der Einfluß

… an Zusammenstellung der Cluster:

- Regelwerk von Bedingungen
  an Vertriebsgebiete und Cluster

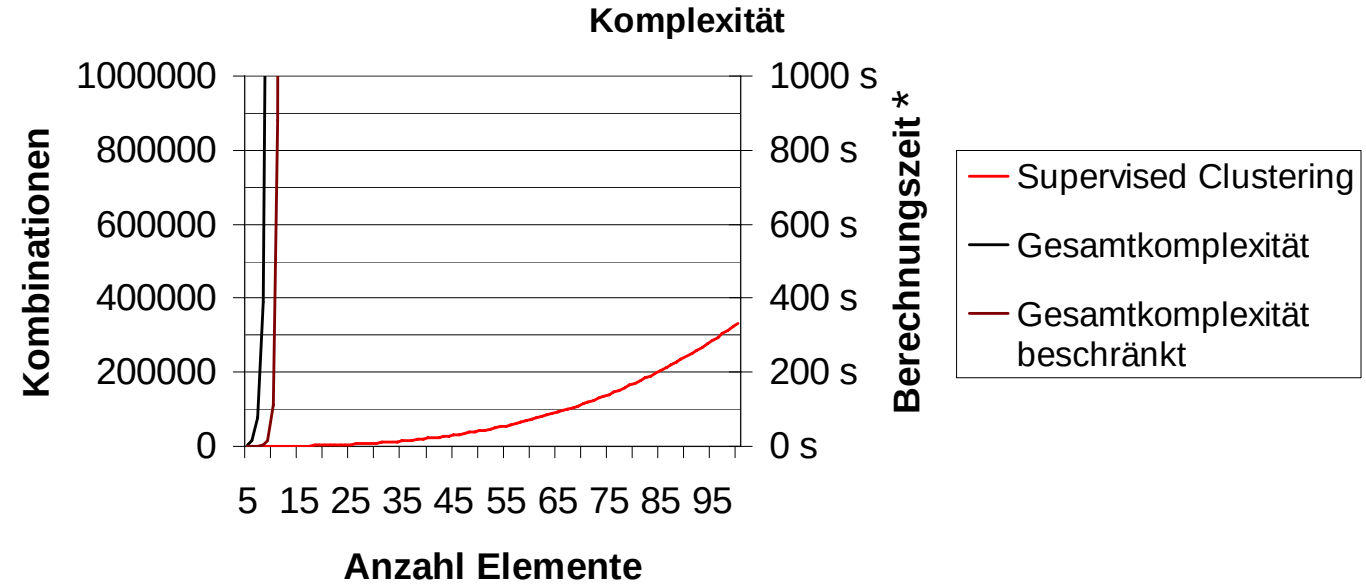
Algorithmus



- Beschränkung auf eine Untermenge der Möglichkeiten
- Abbildung der menschlichen Vorgehensweise:
  - Inkrementelle, lokale Optimierung
  - Ausgehend von Einzelelementen, Elemente zusammenführen, die dadurch eine Verbesserung bzgl. einer Zielfunktion erfahren

Komplexität

- Beispiel: maximal 5 Gruppen mit 50 Elementen
- Anzahl zu untersuchender Möglichkeiten:
  - Gesamtkomplexität: $> 10^{34}$
  - Pro Gruppe genau 10 Elemente: $> 10^{31}$
  - Supervised Clustering: $< 10^{5}$



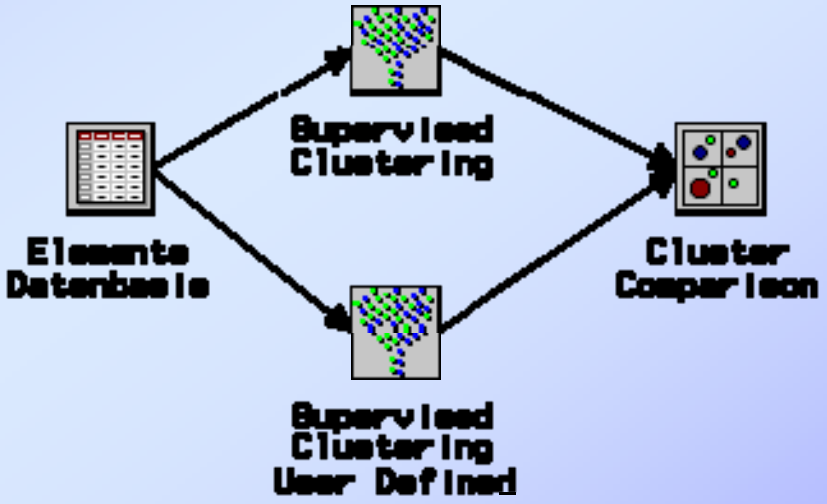
* Ausführungszeit aufgenommen mit Pentium 800 MHz, nichtoptimierter Interpretercode

Implementierung

- Prototyphaftigkeit aufgrund
  - vorher unklarer Anforderungen, Bedingungen
  - Umfang der Verbesserung unbekannt
    (generell und auf Basis spezieller Daten)
- Prototyp in Prolog (logische Programmierung)
  - Datenbasis
    beliebige Daten assoziativ speicherbar
  - Regelbasiert
    einfache nachträgliche Definition von einschränkenden Bedingungen
    bei jeder Aktion
- Stabile Version in SAS
  - Nahtlose Integration in
    - andere Untersuchungen
    - SAS/Enterprise Miner

Implementierung in SAS

- SAS ist insgesamt „datengetrieben“:
  - Steps u. Procs kommunizieren über Datasets u. Macrovariablen
- Algorithmus fest in Data-Steps implementiert

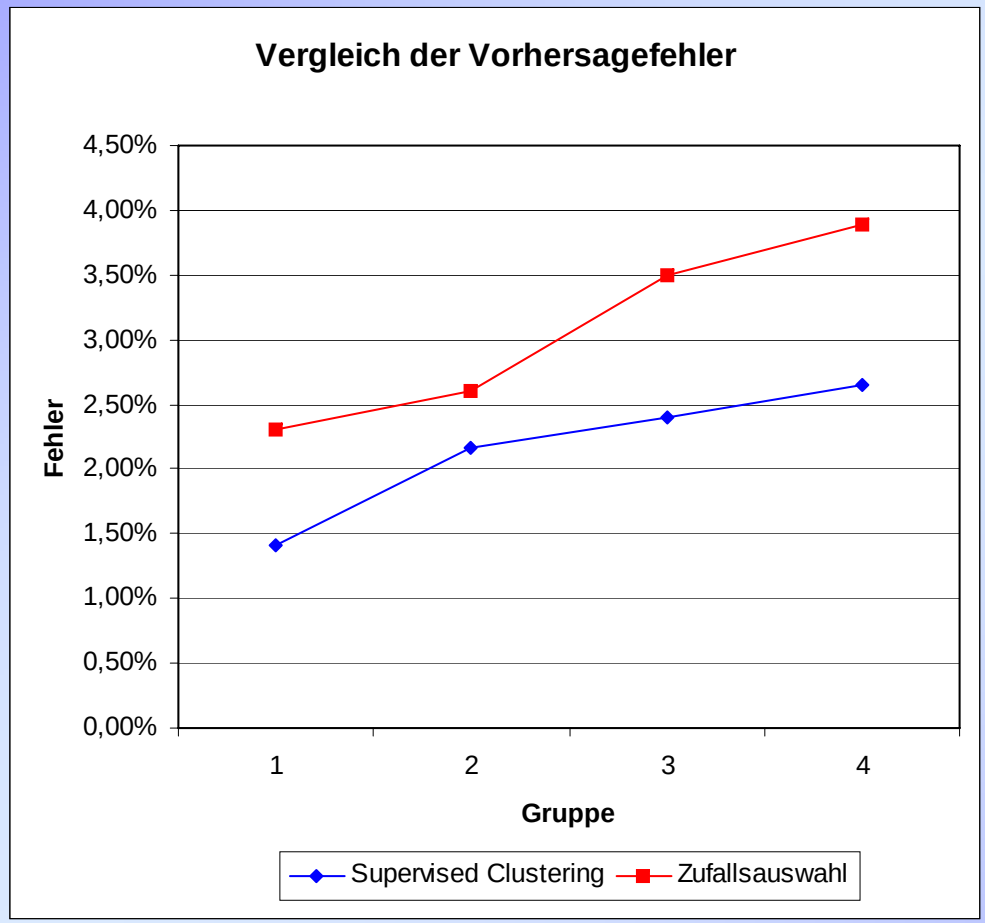


- Extensive Nutzung von SAS/Macro zur funktionalen Wiederverwendung von Data-Steps
- Skalierbarkeit gesichert durch Nutzung

  externer Speicher (Datasets)

  optimierter Dataset-Operationen

Ergebnis

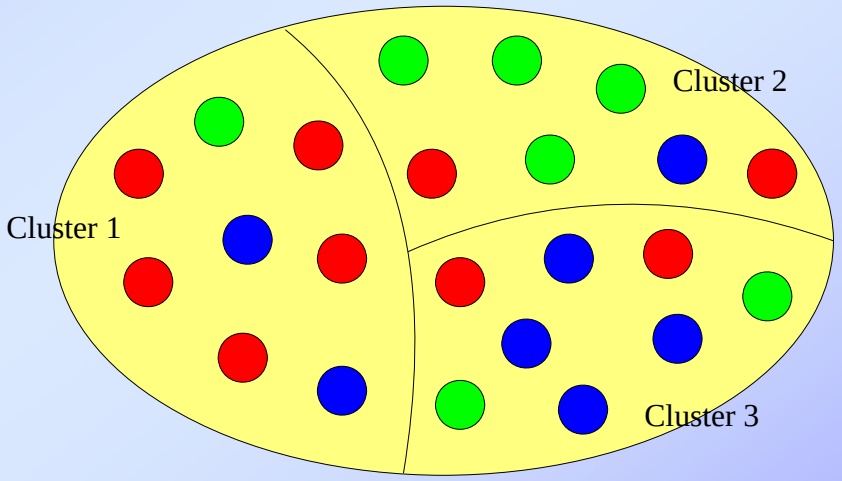


- deutliche Fehler-reduktion in allen Gruppen
- es gibt wenige sich gut für diese Zeitschrift ausgleichende Vertriebsgebiete
  - Volatilität
  - Regionalität

Lösungsvergleich

- Frage:
  - Wie unterscheiden sich verschiedene Gruppierungen?

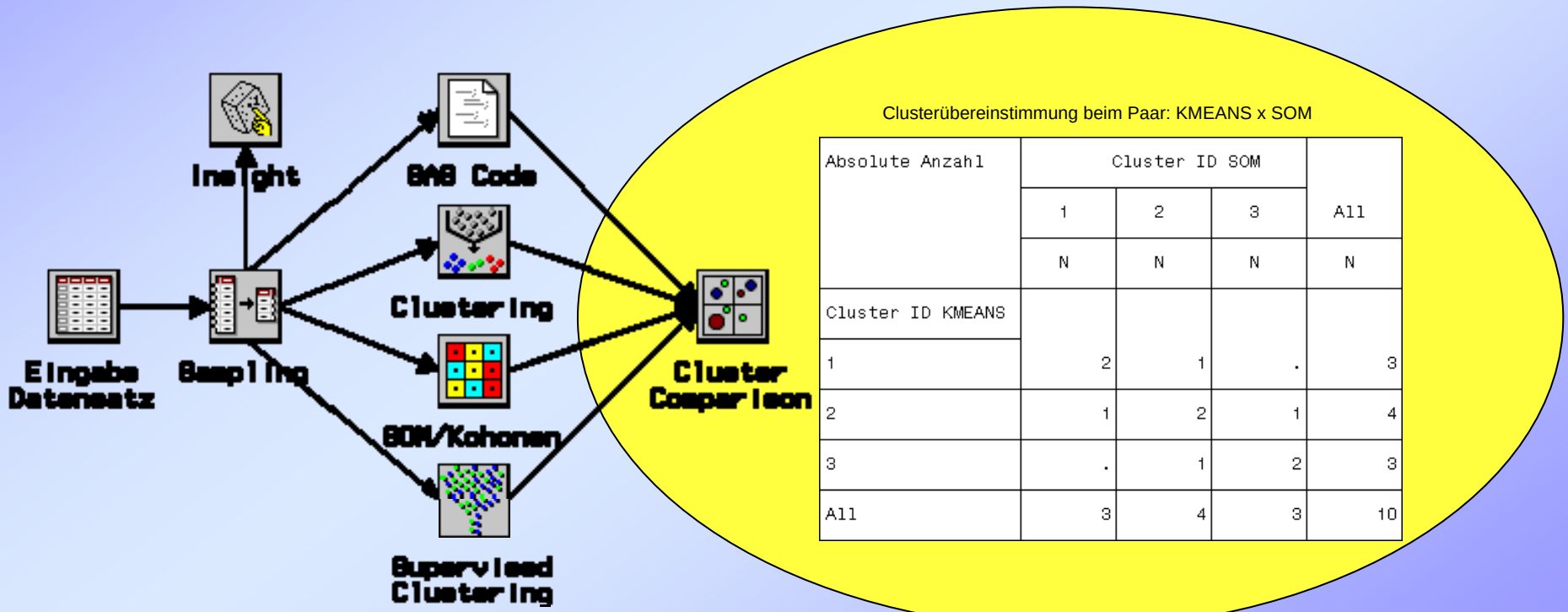
  oder
  - Auf welche Cluster einer anderen Gruppierung verteilen sich die Elemente eines Clusters?



- Mittel: Kreuztabellen

SAS/EM-Knoten: „Cluster Comparison“

- Beliebige Aggregatfunktionen über Elemente möglich: Anzahl, Summe, Min/Max, …

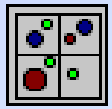


Clusterübereinstimmung beim Paar: KMEANS x SOM

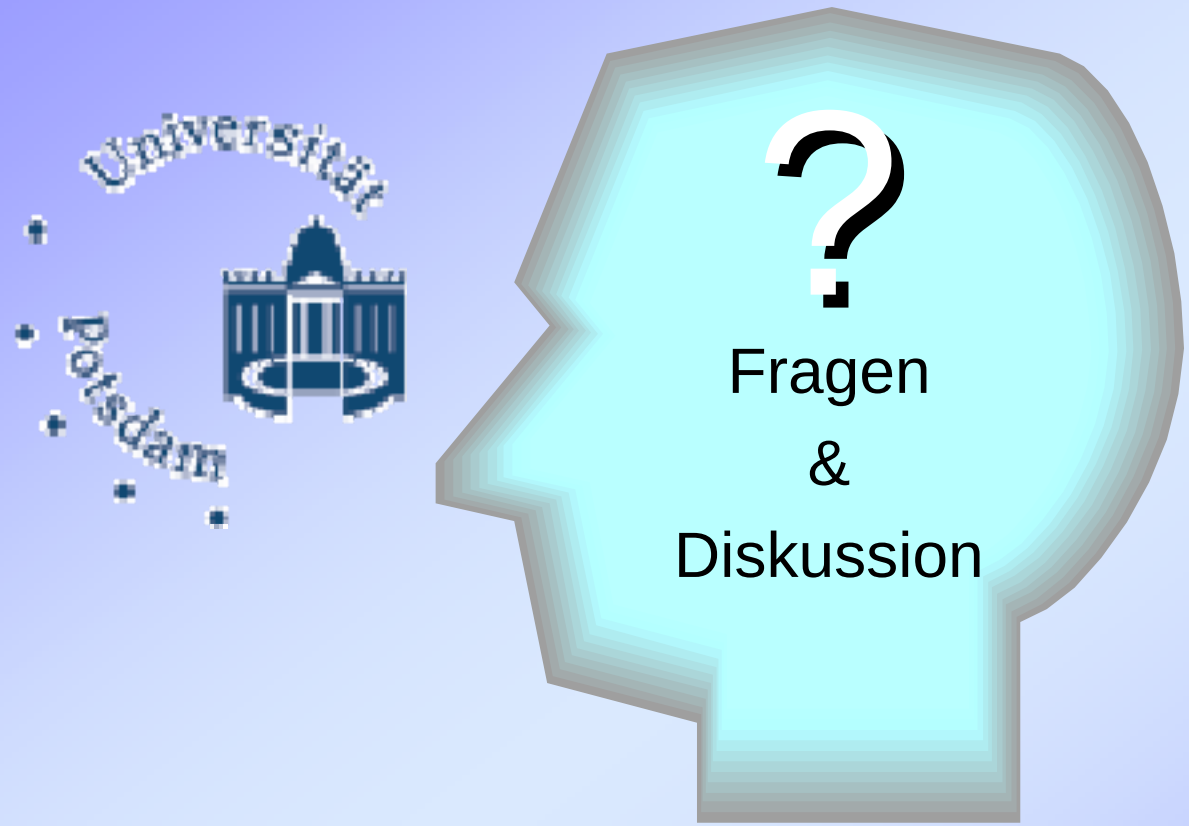
Absolute Anzahl	Cluster ID SOM			
	1	2	3	All
	N	N	N	N
Cluster ID KMEANS				
1	2	1	.	3
2	1	2	1	4
3	.	1	2	3
All	3	4	3	10

Zusammenfassung

- Supervised Clustering
  - Testbarkeit bei auflagenschwachem Titel durch das Verfahren erst möglich geworden
  - Agglomerationsverhalten gibt Auskunft über
    - Eigenschaften der Daten
    - zugrundeliegende Prozesse



- Cluster Comparison
  - Vollständige Integration in Tools des SAS/EM
  - Überlappungsgrad gibt Hinweis auf Stabilität einer Aggregation



KONFERENZ DER SAS®-ANWENDER IN FORSCHUNG UND ENTWICKLUNG

VERLAGVERTRIEBS KG

Stefan Pohl: mail@s-pohl.de
Dr. Sergej Steinberg: s.steinberg@bauerverlag.de