



# Datenvorverarbeitung von nominalen Daten für Data Mining

Entstanden 2004/2005 bei der  
T-Systems International GmbH  
unter Betreuung von  
Prof. Dr. J. Fürnkranz



# Gliederung

- Datenvorverarbeitung
- Prepared Information Environment
- Vorverarbeitung von nominalen Daten
- Exponieren semantischer Information

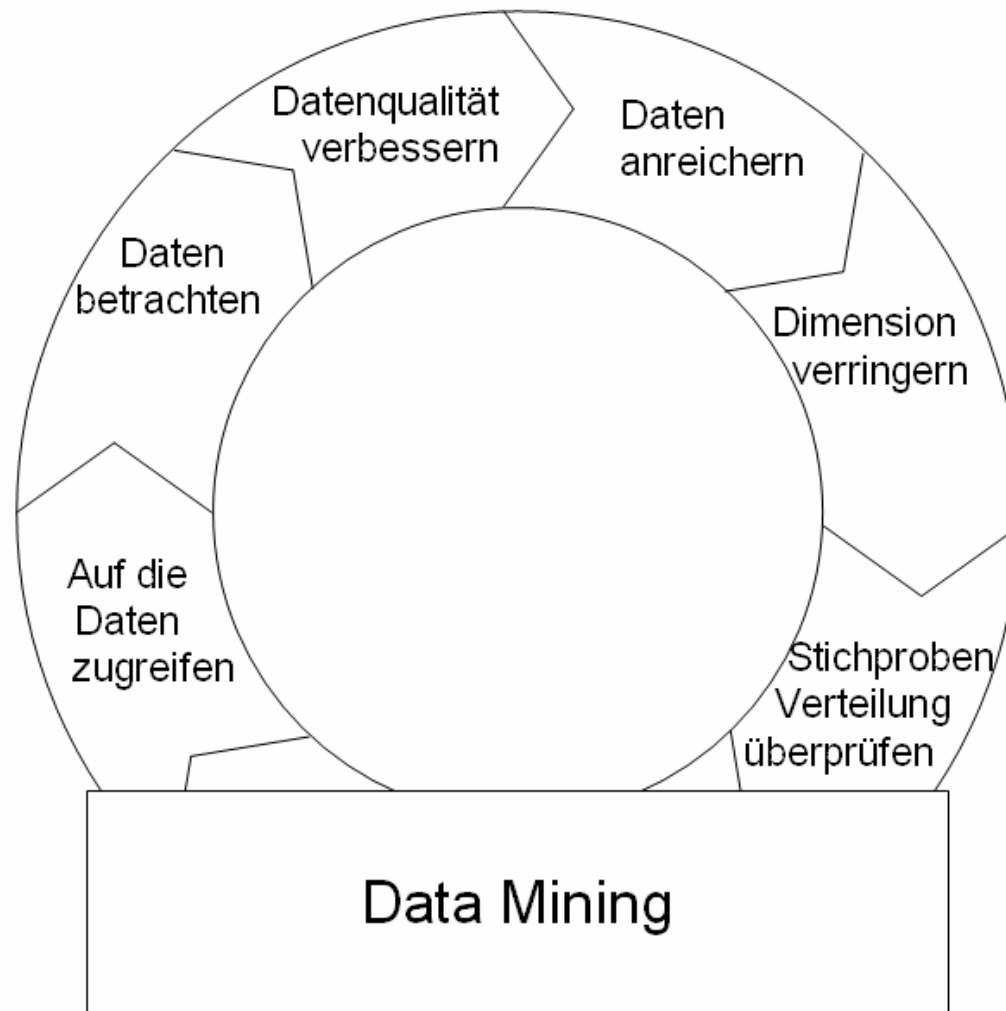


# Datenvorverarbeitung



# Datenvorverarbeitung

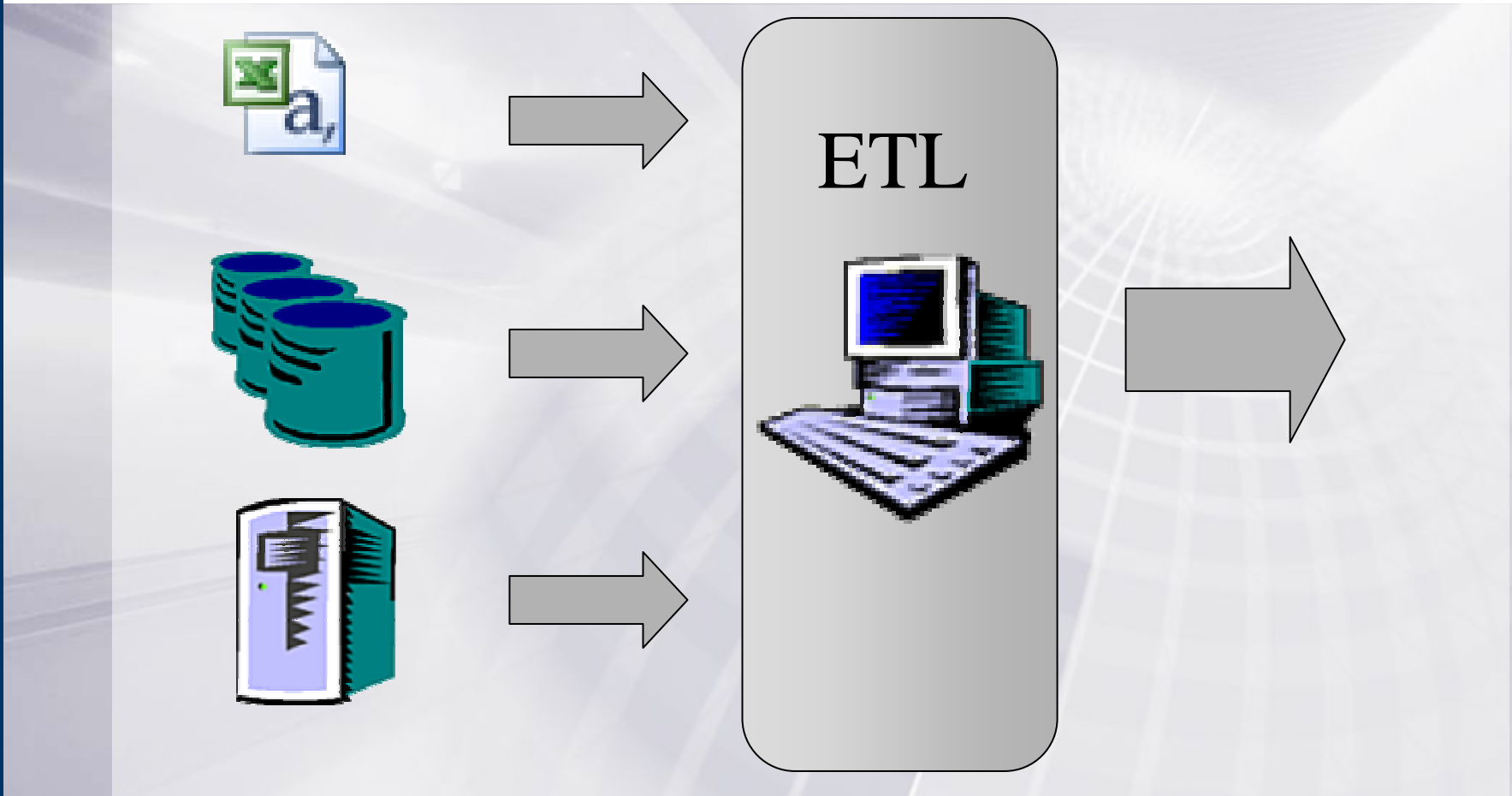
## Der Zyklus





# Datenvorverarbeitung

## Auf die Daten zugreifen





# Datenvorverarbeitung

## Daten betrachten

- Attribute verstehen
- Fehler erkennen
- Erste Hypothesen



Umgesetzt mit

- OLAP Tools
- SQL

SQL



QUEST  
SOFTWARE



# Datenvorverarbeitung

## Datenqualität verbessern

- Behandlung von Ausreißern
- Ungültige Werte
- Falsche Formate
- Dubletten
- Fehlende Werte
- Widersprüchliche Werte





# Datenvorverarbeitung

## Daten anreichern

### Ergänzen um weitere Attribute

- Aggregationen von Werten
- Attribute aus externen Quellen
- Einbeziehen von Expertenwissen



Domäne des  
menschlichen Experten

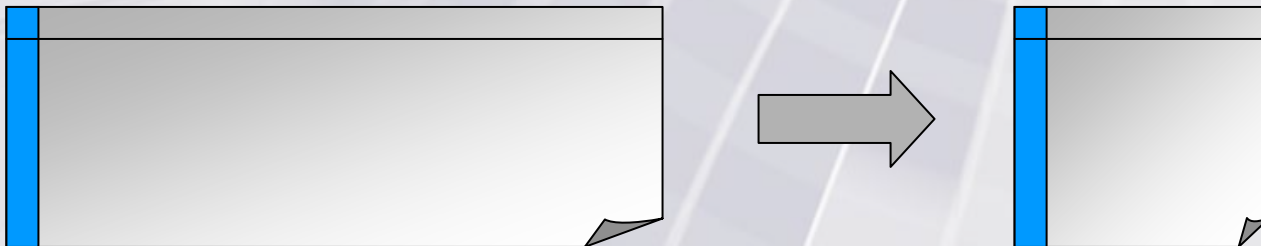




# Datenvorverarbeitung

## Dimension verringern

- Vermeiden von Overfitting
- Reduzieren der Rechenzeit
- Entfernen von Selbstbezügen
- Entfernen von Attributen, die in der Praxis nicht zur Verfügung stehen

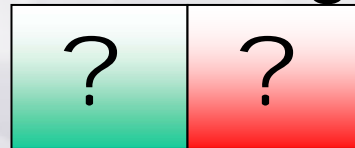




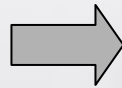
# Datenvorverarbeitung

## Stichprobenverteilung prüfen

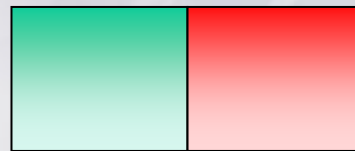
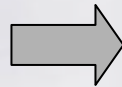
- Zufällige Verteilung des Targets



- Erhaltung der Verteilung des Targets



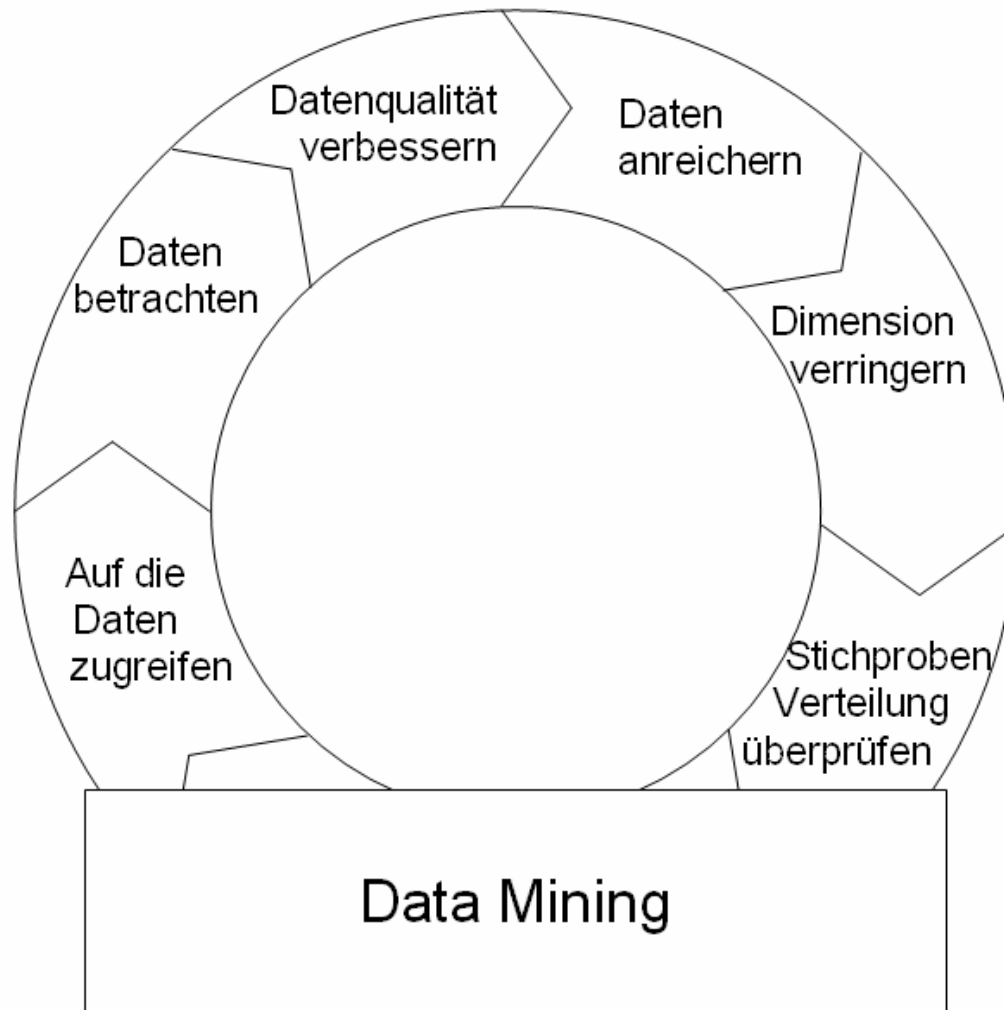
- Gleichmässige Verteilung von Target





# Datenvorverarbeitung

## Der Zyklus



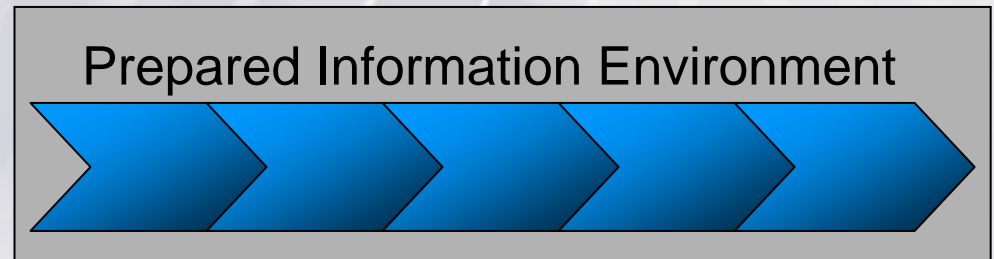


# Prepared Information Environment



# Prepared Information Environment Überblick

- Ein Rahmen für die Vorverarbeitung
- Vereinfacht das erneute Anstoßen des Zyklus
- Sichert die Replizierbarkeit des Modells
- Routine bei periodisch zu erstellenden Modellen
- Spart Zeit





# Prepared Information Environment

## PIE mit SQL

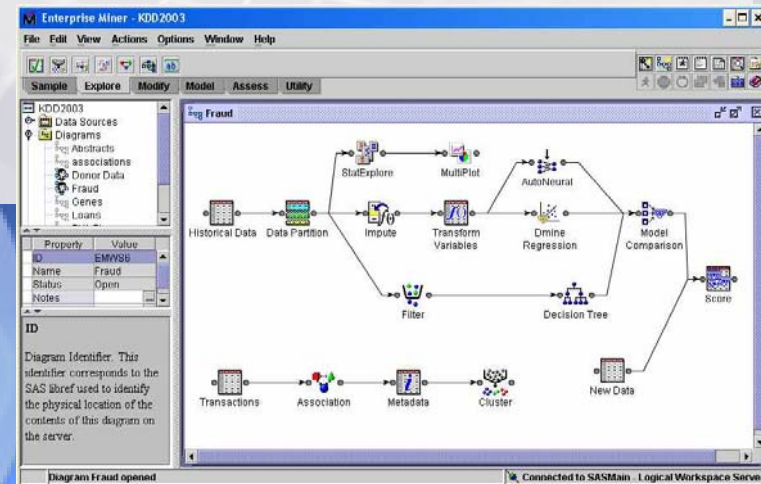
- Transformationen lassen sich mit SQL durchführen
- SQL Skripte können automatisiert erneut aufgerufen werden
- Ist bereits ein Nebenprodukt des Datenvorverarbeitungsprozesses





# Prepared Information Environment PIE mit SAS

- Prozesse in einem Flussdiagramm.
- SAS 9 ist konzipiert um alle Aufgaben der Datenvorverarbeitung durchzuführen
- Bedingt transportabel





# Prepared Information Environment

## Hybrides PIE

- Zusammenfügen von SQL-Skripten und SAS 8.2
- Output über ID anstelle eines Output-Moduls
- Flexibel
- Geringer Mehraufwand







# Normalisieren von Schreibweisen



# Wissen in der Datenvorverarbeitung

- Unstrukturierte Felder enthalten menschenlesbare Informationen
  - Produktnamen, Titel, Kommentare, Namen,...
- Anreichern der Daten ist eine kreative Arbeit
- Erfahrung und Fachwissen wird für viele Entscheidungen benötigt

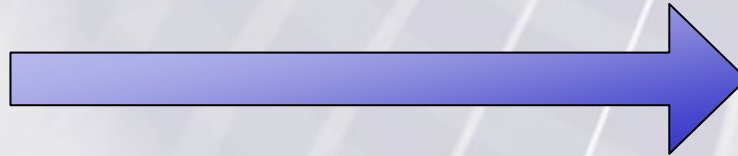




# Wissen in der Datenvorverarbeitung

## Datenbanken

- Expertenwissen zum befüllen notwendig
- Auflistung existenter Nachnamen
  - Befüllung sehr aufwändig
  - Datenbank wird sehr gross
- Angereichert mit Metadaten





# Wissen in der Datenvorverarbeitung

## Heuristiken

- Levenshtein Distanz

- Präfix Matching

- Jaro Ähnlichkeitsmaß

$$\text{Jaro}(s;t) = \frac{1}{3} \cdot \left( \frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s'|} \right)$$

- Jaro-Winkler Maß

$$\text{Jaro-Winkler}(s; t) = \text{Jaro}(s; t) + \frac{P'}{10} \cdot (1 - \text{Jaro}(s; t))$$

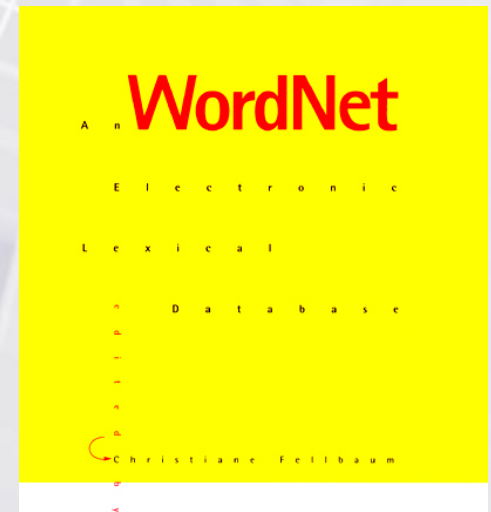
- Smith-Waterman Algorithmus



# Wissen in der Datenvorverarbeitung

## Semantische Netzwerke

- Knoten beinhalten Begriffe
- Kanten repräsentieren Relationen
- Semantische Distanz zwischen Knoten
- Geeignet um Zusammenhänge und Assoziationen zu verarbeiten
  - Homonyme
  - Synonyme
  - Antonyme
  - Hyperonyme
  - Hyponyme
  - Meronyme

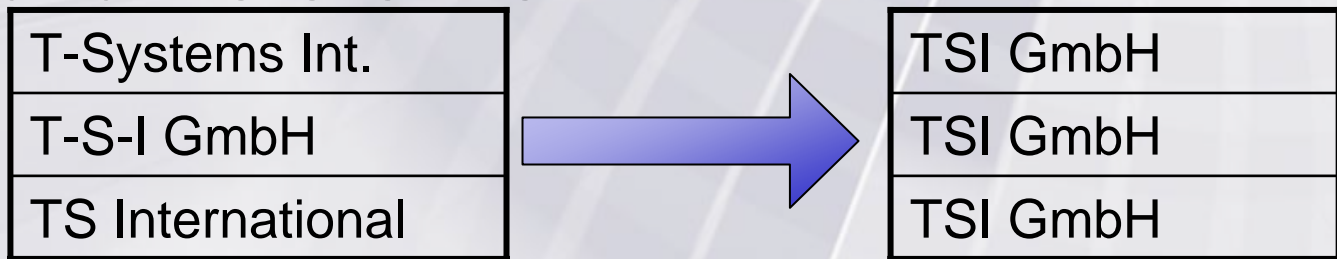




# Wissen in der Datenvorverarbeitung

## Regeldatenbanken

- Regeln können wiederkehrende Fehler beheben
- Regeln können Wortformen normalisieren
- Eine Regeldatenbank enthält Regeln und Referenzen





# Normalisieren von Schreibweisen

## Problemstellung

- Welche Zeichen sind falsch?
- Welche Zeichen fehlen?
- Welche Zeichen sind verdreht?
- Welche Worte sind falsch?
- Welche Worte sind korrekt, gehören aber woanders hin?
- Welche Worte können unterschiedlich geschrieben werden?





# Normalisieren von Schreibweisen

## Lösungskonzept

- Eindeutige Form notwendig
- Anwendung von Regeln
- In unbekannten Fällen werden mit Hilfe des Experten weitere Regeln erstellt
- Datenbank mit korrekten Formen
- Vergleich von Formen über Heuristiken

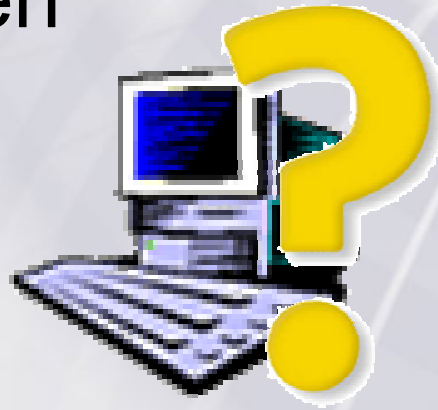




# Erkennen von semantischen Fehlern

## Problemstellung

- Sind automatisiert fast nicht zu erkennen
- Der Computer kann die menschenlesbaren Informationen nicht ohne Hilfe verarbeiten





# Erkennen von semantischen Fehlern

## Lösungskonzept

- Sichtung durch einen Experten
- Erkennen des Fehlers am Ende eines Datenvorverarbeitungszyklus
- Zur Automatisierung ist Wissen erforderlich
- Externe Informationen müssten automatisch herangezogen werden.

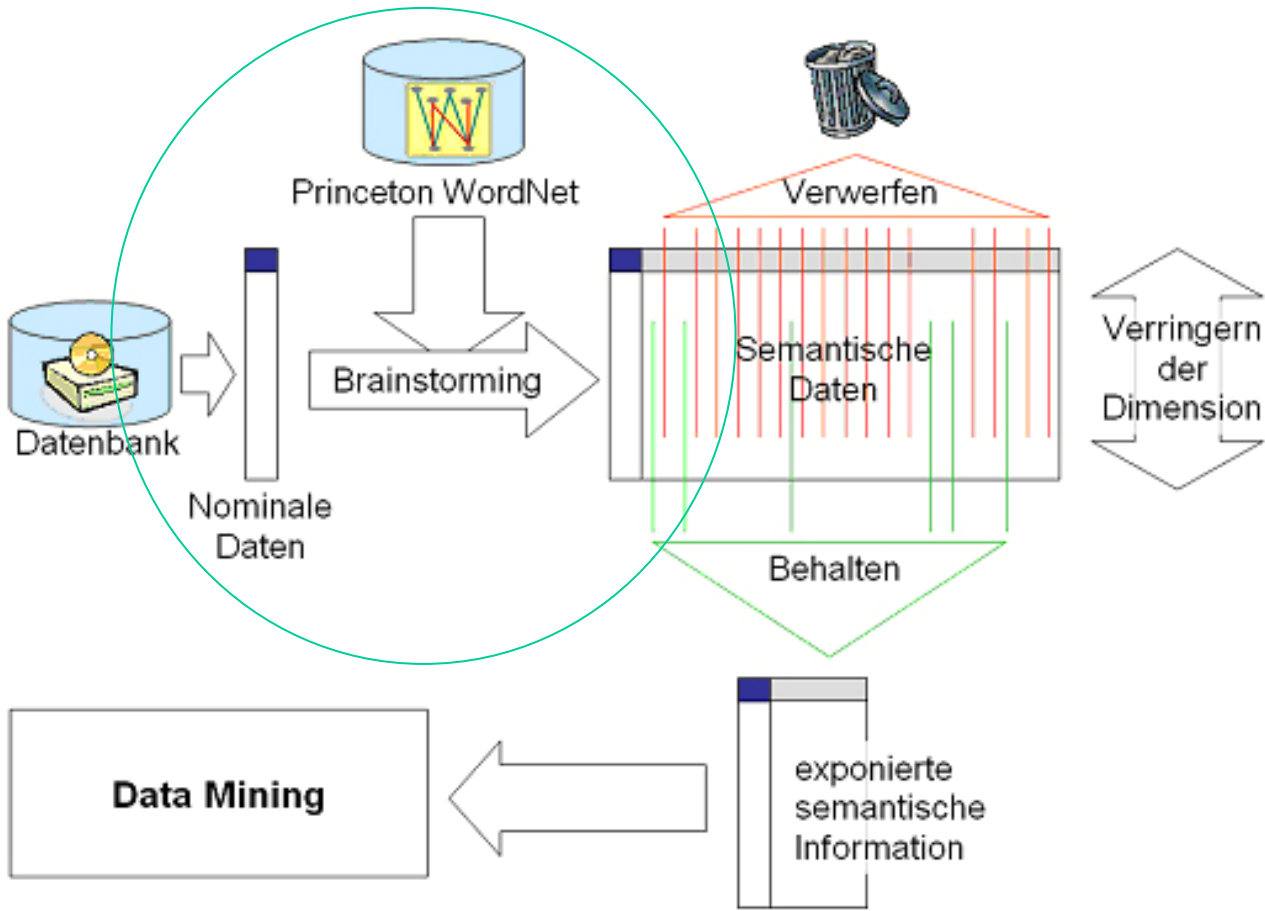




# Exponieren semantischer Information

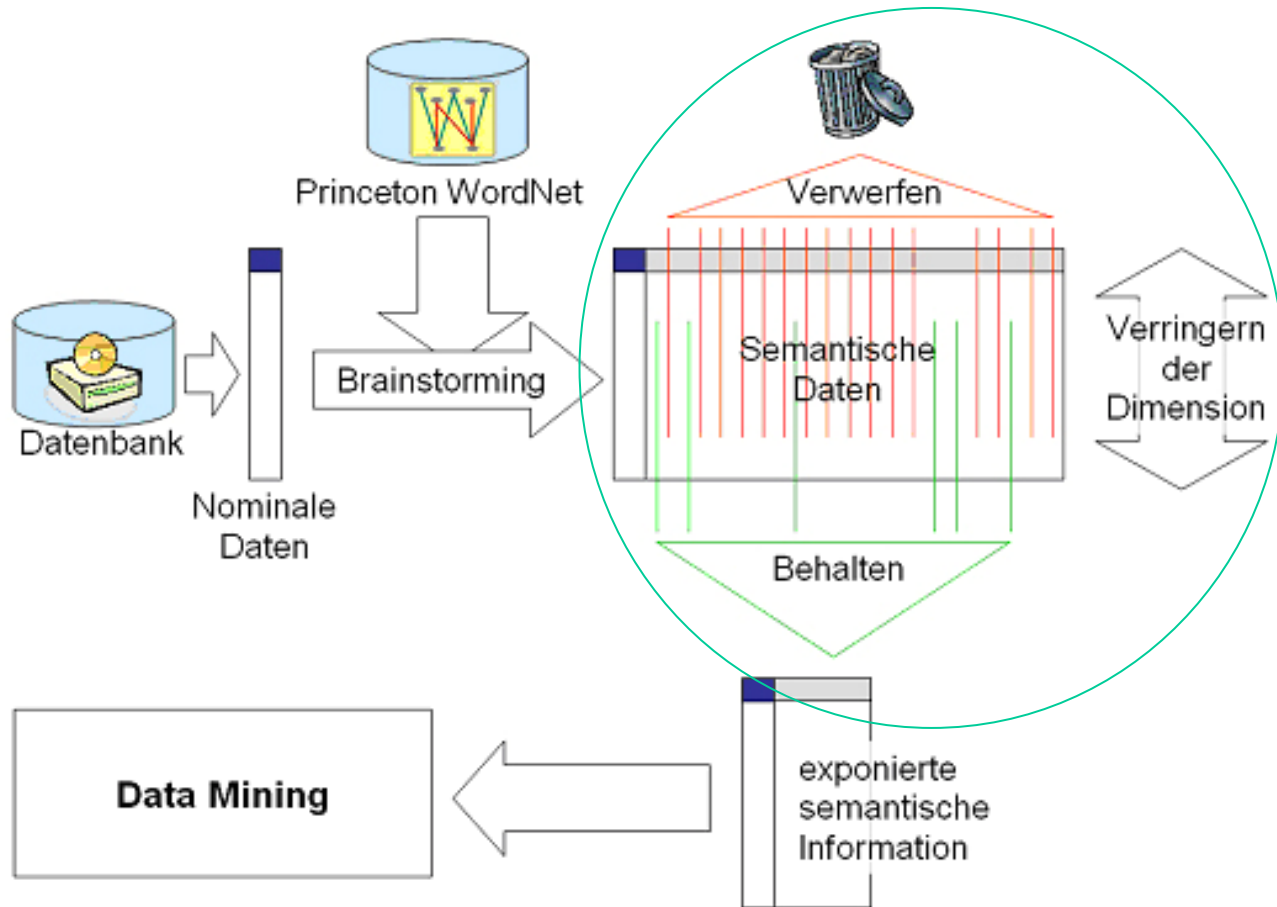


# Exponieren semantischer Information Brainstorming





# Exponieren semantischer Information Verringern der Dimension





# Exponieren semantischer Information

## SCM Tool

Semantic Content Mining

File Options Help

Content	quality	design	activity	condition	status	region
Ben-Hur (1959)	0.06	0.13	0.13	0.06	0.06	0.25
Ninotchka (1939)	.	.	.	.	.	.
Good Will Hunting (1997)	1.00	1.00	1.00	0.25	0.25	0.13
Silence of the Lambs, The (1991)	0.13	0.13	0.25	1.00	1.00	0.06
Wonderland (1997)	0.25	0.25	0.25	0.25	0.25	0.50
Nikita (La Femme Nikita) (1990)	.	.	.	.	.	.
Field of Dreams (1989)	0.25	0.25	0.50	0.25	0.25	1.00
Pulp Fiction (1994)	0.13	0.25	0.25	0.25	0.13	0.13
Apostle, The (1997)	.	0.02	.	.	.	0.02
Blues Brothers, The (1980)	0.13	0.06	.	.	.	0.13
African Queen, The (1951)	0.25	0.25	.	.	.	0.13
Flubber (1997)	.	.	.	.	.	.
Star Trek III: The Search for Spock ...	0.50	0.25	.	.	.	0.25
My Fellow Americans (1996)	0.25	0.06	.	.	.	0.13
Mr. Magoo (1997)	0.25	0.06	.	.	.	0.13
Rising Sun (1993)	0.25	0.25	.	.	.	0.25
Mulholland Falls (1996)	0.06	0.02	.	.	.	0.03
Dumb & Dumber (1994)	.	.	.	.	.	.
Cinderella (1950)	0.50	0.13	.	.	.	0.25
Amityville: Dollhouse (1996)	0.13	0.13	.	.	.	0.50
Mask, The (1994)	.	0.02	.	.	.	0.01
Return of the Jedi (1983)	0.50	0.25	.	.	.	0.13
Long Kiss Goodnight, The (1996)	0.13	0.13	.	.	.	0.06
Hearts and Minds (1996)	.	.	.	.	.	.
Terminator, The (1984)	0.25	0.06	.	.	.	.
Poetic Justice (1993)	0.25	0.13	.	.	.	0.13
Independence Day (ID4) (1996)	0.13	0.13	.	.	.	0.13
FairyTale: A True Story (1997)	0.13	0.25	.	.	.	0.25

Ready

**Options**

Load

First Row is Caption

Expose Information

Include Tokenizer

Expose Antonyms

Expose Hyponyms

Extend exposed Information

1 Automated Steps

Extend with Antonyms

Extend with Hyponyms

Reduce Dimension

15 Minimal percent for keeping attributes

15 Minimal percent for keeping duplicates

10 Minimal number of attributes kept.

100 Maximal number of attributes kept.

Shrink table by Entropy.

Help OK Cancel

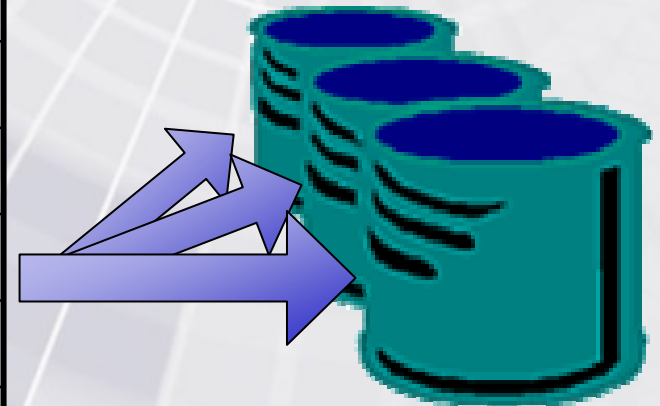


# Exponieren semantischer Information

## Evaluationszenario

- Verwendung der offiziellen MovieLens Datenbank
- Prognose der Bewertung anhand des Kinotitels
- Angereichert um manuell erstellte Genre
- Angereichert um mit SCM erstellte Attribute

Up in Smoke (1978)
Two Deaths (1995)
Safe Passage (1994)
Nine Months (1995)
Money Train (1995)
...





# Exponieren semantischer Information

## Evaluationsresultat

Daten	Wahr-positiv	Gesamter Fehler	Normierte Präzision
Genre	0,0%	21,3%	-1,00
SCM 2.0	8,8%	22,3%	-0,82
SCM 2.5	14,7%	22,9%	-0,70
Genre_SCM 2.0	8,8%	23,9%	-0,82
Genre_SCM 2.5	5,8%	21,3%	-0,88





Vielen Dank für ihre Aufmerksamkeit

Oliver.Werth@Athistaur.de