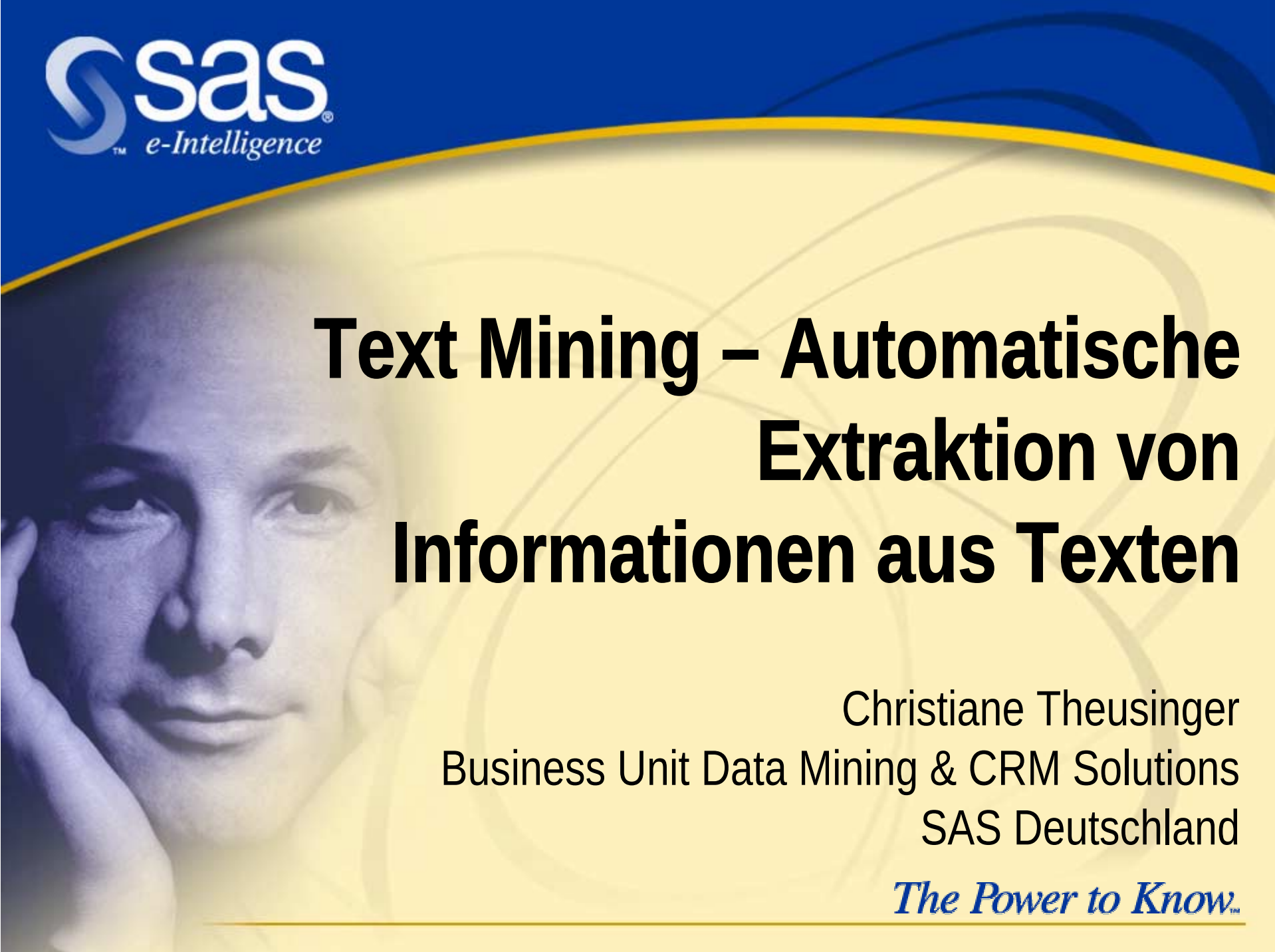


# Text Mining – Automatische Extraktion von Informationen aus Texten

Christiane Theusinger
Business Unit Data Mining & CRM Solutions
SAS Deutschland

The Power to Know.

# Agenda

- Einführung Text Mining
- Übersicht über Anwendungen für Text Mining
- Die Text Mining-Lösung von SAS

# Definition Text Mining

Knowledge

Entdeckung von unbekanntem und potentiell nützlichem Wissen aus Texten bzw. Textsammlungen

# Warum Text Mining ?

- Immer mehr Informationen liegen in digitaler Form vor
- Fülle an Informationen fast nicht verarbeitbar
- Konkurrenzdruck: schnell und kostengünstig reagieren

# Ziele Text Mining

- **Klassifikation**

  eindeutige oder sich überlappende Kategorien

- **Segmentierung**

  aufgrund von Ähnlichkeiten

- **Navigation**

  inhaltlich zusammenhängende Dokumente

# Text Mining Anwendungen

## Automatische Klassifikation von Dokumenten

- Automatisches Filtern von Emails

## Segmentierung

- Aufbau einer Wissensbasis
- Themen und Konzepte identifizieren

## Vorhersage

- Kategorisierung von Hotline-Anrufen

## Ergänzung zu Data Mining-Analysen

- Integration von Textinformationen in Analysen

# Help Desk

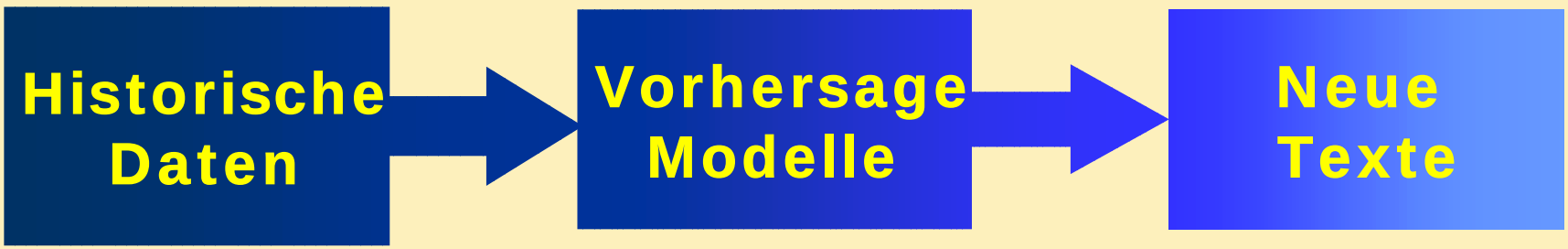
| | | |
|---|---|---|
| Fälle | = | Problembeschreibung |
| Input Variablen | = | Wörter, Konzepte, Phrasen |
| Zielvariable | = | Response auf Problem |
| Aktion | = | automatisch Hilfe /Unterstützung bereitstellen oder Problem an Spezialisten weiterleiten |

# Bearbeitung von Kundenbriefen

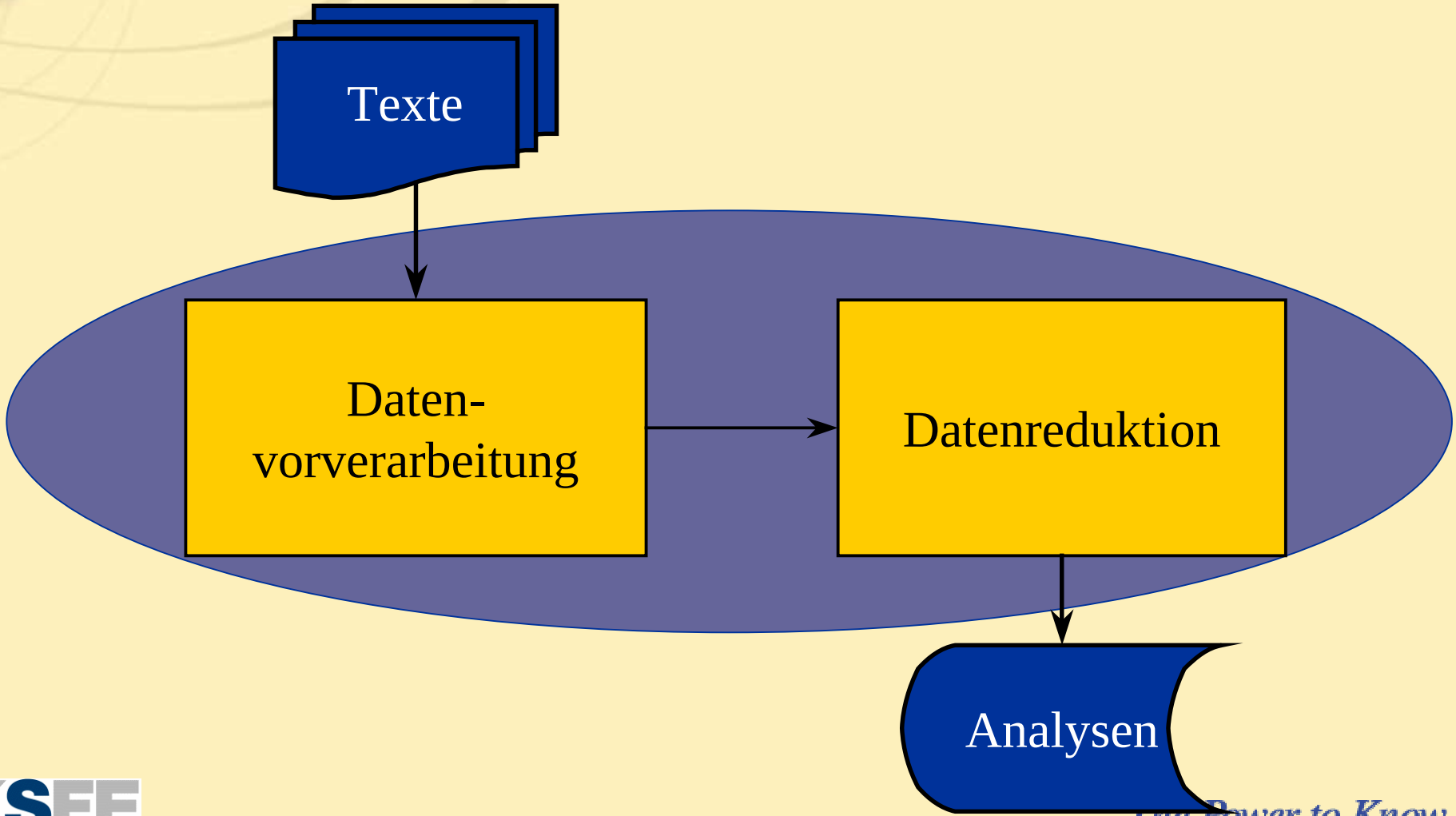
| | | |
|---|---|---|
| Fälle | = | Eingescannte oder digitale Kundenbriefe |
| Input Variablen | = | Wörter, Konzepte, Phrasen |
| Zielvariable | = | Response auf Problem |
| Aktion | = | Automatische Response (z.B. Aussendung von Informations-material) oder Weiterleiten des Briefes an zuständige Abteilung |

# Text Mining Vorgehensweise

Nutze die Daten von durchgeführten Klassifikationen, um neue Texte zu klassifizieren



# Die SAS Text Mining-Lösung



# Schritte des Text Minings

- Texte einlesen
- Datenvorverarbeitung
- Dimensionsreduktion
- Text Mining zur Klassifikation nach Themen
  - Einsatz verschiedener Verfahren (u.a. Neuronale Netze, Regression )
  - Hinzufügen weiterer Variablen zu den Ergebnissen

# Datenvorverarbeitung

- Einlesen des Textes
- Analyse der Satzstruktur
- Abgleich mit bekannten Mustern (z.B. Personennamen, Firmennamen)
- Analyse der Wortmorphologie und Eliminierung von irrelevanten Wörtern (u.a. Artikel)
- Erstellen einer Häufigkeitstabelle der Ausdrücke

# Stoppliste

Eliminierung von Wörtern,die keine sinnvolle eigenständige Bedeutung haben, wie  Artikel, Konjunktionen oder Präpositionen

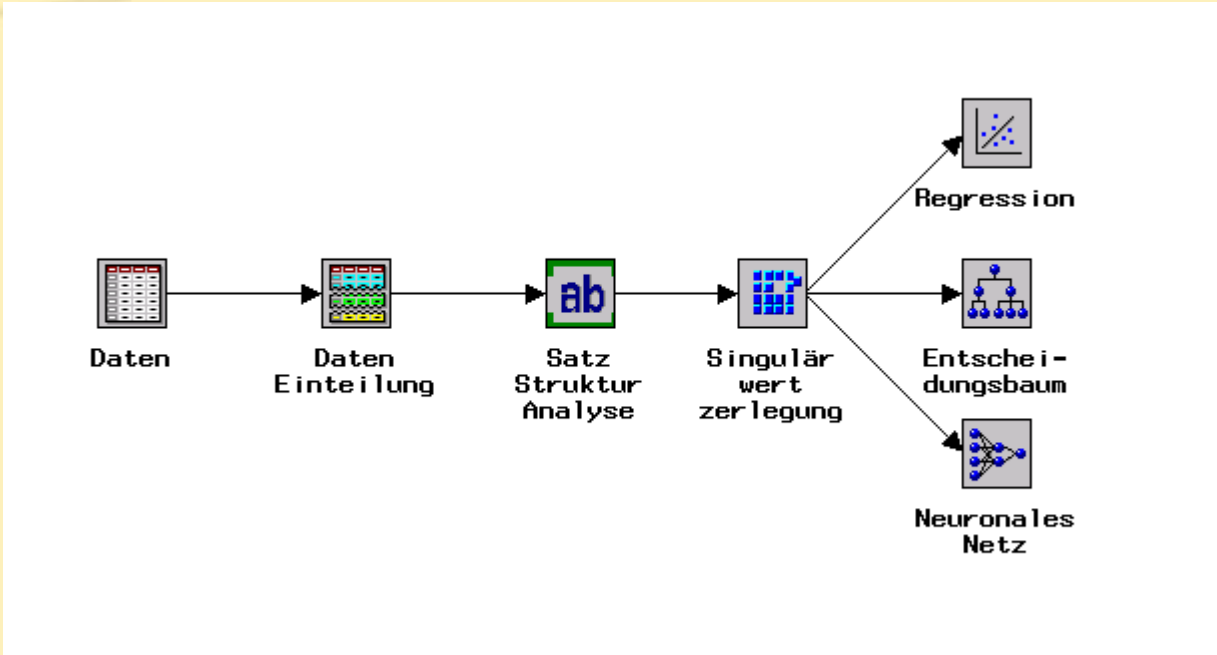
# Darstellung

| | Dokument | | |
|---|---|---|---|
| | 1 | 0 | 2 |
| **Wort** | 0 | 1 | 1 |
| | 3 | 0 | 0 |

# Dimensionsreduktion

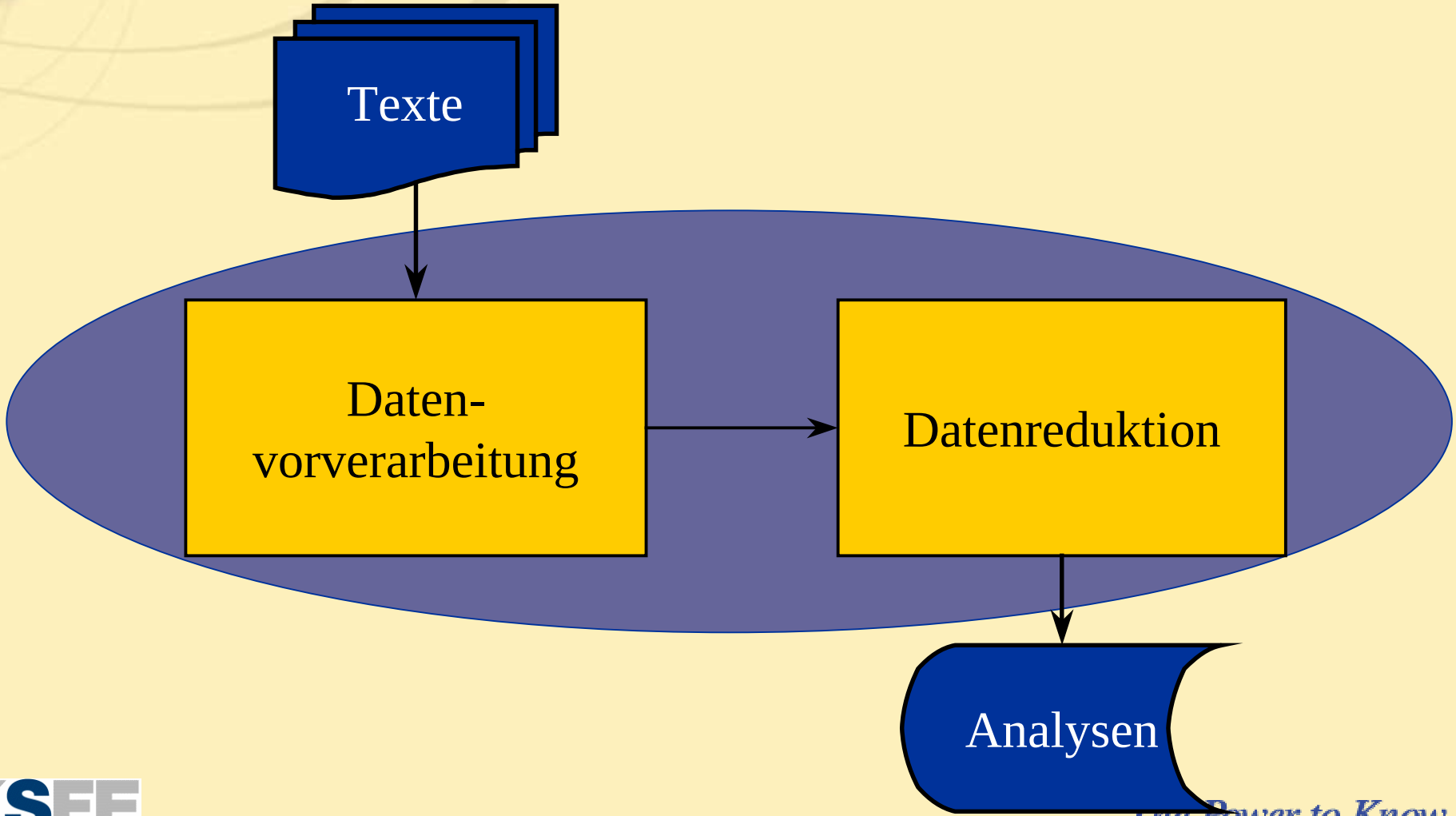
- Reduktion auf ein- oder zweihundert Dimensionen
- Ermöglicht Abschätzung, wann zwei Dokumente ähnlich sind
- Eingesetzes Verfahren: Singulärwertzerlegung

# Text Mining Prozessfluss



# Kategorisierung von Nachrichtentexten

# Die SAS Text Mining-Lösung



# Zusammenfassung

- SAS hat ein Text Mining-Add-on zum Enterprise Miner entwickelt.
- Integrierte Lösung für Text und Data Mining
- Status: Experimentell mit SAS 8.2/EM 4.1

SAS Deutschland
In der Neckarhelle 162
69118 Heidelberg

*Tel.:* 06221/ 41 5 0
*Fax:* 06221/ 41 51 01

*Internet:* www.sas.com
*Email:* Christiane.Theusinger@ger.sas.com