

# TNS EX·A·MINE™

## TextMiner

Aufdecken von Beziehungen und  
Trends in unstrukturierten Daten





- In den letzten Jahren hat sich der **Information Overload** vervielfacht
- Informationen in **unstrukturierter und halbstrukturierter Form**
- Anteil von unstrukturierten Informationen: **80 bis 90 %**
- Wie können diese Informationen erfasst, untersucht und genutzt werden?
- **Text Mining** ist die Analyse von gesammeltem Textmaterial
- **Ziele** des Text Mining:
  - **Schlüsselbegriffe und Themen** erfassen
  - **Beziehungen und Trends** aufdecken
- Text Mining bietet eine **automatisierte Inhaltsanalyse**, basierend auf einer Kombination **linguistischer** und **statistischer Methoden**



## ■ **Marktforschung**

- Ermittlung von zentralen Themen bei offenen Fragen
- Was wird positiv bewertet, was negativ?
- Welche Begriffe werden kombiniert genannt?
- Welche übergeordneten Kategorien gibt es?

## ■ **Customer Relationship Management (CRM)**

- Berücksichtigung aller Kundenberührungspunkte (E-Mails, Call Center Kontakte, Transaktionen, Umfragen, ...)
- Welches sind die Hauptanliegen und -probleme der Kunden?
- Welche Beschwerden und Problemkonstellationen treten gemeinsam auf?

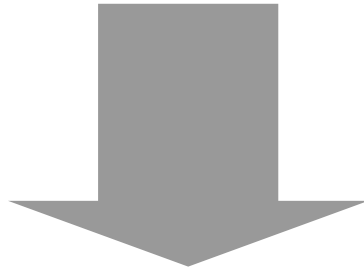
## ■ **Analyse von Blogs, Foren und Web-Feeds**

- Untersuchung der zentralen Schlüsselbegriffe aus News-Feeds, Blogs, Foren usw.
- Auf welche Themenkategorien lassen sich Beiträge zusammenfassen?
- Konkurrenzbeobachtung und Ideengenerierung mit Hilfe von Blogs

## ■ **Betrugserkennung, medizinische Forschung, ...**



## Identifikation von Konzepten



## Kategorisierung



## Definition von Oberkategorien



## Zielgruppenanalyse

- Konzept = Wort oder Wortkombination
- Linguistisch basierte Auswertung
  - Fokus auf Wörter mit Aussagegehalt (Substantive, Verben, Adjektive, Adverbien, ...)
  - Systematische Ausklammerung von inhaltlich nicht zielführenden Wörtern wie Artikel, Präpositionen, Pronomen etc.
  - Berücksichtigung von Beugungen der Wortstämme (Konjugation/Deklinationen, ...)
- Bündelung von inhaltlich gleich gerichteten Konzepten zu Kategorien
- Weitere Bündelung von Kategorien zu Oberkategorien
- Analyse interessierender Untergruppen mittels deskriptiver, bi- und multivariater Verfahren, z.B. Klassifikation, Assoziationsanalyse, Clusteranalyse, Treiberanalyse



## Algorithmen, die auf die Rohkonzepte angewendet werden:

- Inflection (*vasopeptidase inhibitors = vasopeptidase inhibitor*)
- Synonymy
  - Full-Form: an entire extraction is equivalent to another (*familial hyperchylomicronemia = familial lipoprotein lipase deficiency*)
  - Component: two distinct extractions are equivalent, modulo variation in components (*colour blindness = color blindness*)
- Omission of keywords (*ziff-davis inc = ziff davis*)
- Geographic variant (*tumour = tumor*)
- Lexical variant (*geographical markets = geographic markets*)
- Lower-case/upper-case characters (*apolipoprotein A = apolipoprotein a*)
- Omission of/variation in function words (*ulceration of the mucosa = ulceration of mucosa*)
- Variants in separators; Separators may be space, hyphen, agglutination, apostrophe, or dot (*zollinger-ellison syndrome = zollinger ellison syndrome; health care = healthcare; web-tv = web tv*)
- Inversion of components (*generalized myotonia of Becker = Becker's generalized myotonia; cancer of the thyroid = thyroid cancer; zeste râpé d'un citron = zeste de citron râpé*)
- Accented/non-accented characters -very frequent in languages such as French, Spanish, Italian, or Dutch (*saõ Paulo = sao Paulo; evguéni primakov = evgueni primakov*)
- Generic-specific; Grouping extracts under a normalized term can be seen as finding the “best descriptor.” In some applications, specific terms could be mapped to generic terms (*lipstick = cosmetics*)
- Spell checking/fuzzy matching based on omission of vowels or double consonants, or other algorithms (*technical support = technical support; techinical support = technical support*)

# Kategorienmodell

## Klassifikation von Datensätzen / Dokumenten



**(fiktive Daten)**

Res...	Q1_What_do_you_like_most_about_this_portable_music_player	... REF1_Pr...	...	...	...	...	volume	mobility	software	hardware	...
1	little, light	... Other	...	...	...	...	0	0	0	0	0
2	The battery power is great.	... Product E	...	...	...	...	0	0	0	0	0
3	cost and size	... Other	...	...	...	...	0	0	0	0	1
4	Having all my CDs in the palm of my hand!	... Product A	...	...	...	...	0	0	0	0	0
5	The shuffle mode.	... Product A	...	...	...	...	0	0	0	0	0
6	Battery life. Portability. Accessories. Style.	... Product A	...	...	...	...	0	1	0	0	0
7	I like its ability to store all of my music. I also like the ability to create playlists.	l... Product A	...	...	...	...	0	0	0	1	0
8	portability, capacity, sound quality, durability	i... Other	...	...	...	...	0	1	0	0	0
9	Small, great sound, capacity.	... Product A	...	...	...	...	0	0	0	0	0
10	Able to hold all of my songs in one place.	l... Product A	...	...	...	...	0	0	0	0	0
11	It's portable! I can take it anywhere.	... Product A	...	...	...	...	0	0	0	0	0
12	Living in my own little world	... Product A	...	...	...	...	0	0	0	0	0
13	mobility	... Product A	...	...	...	...	0	1	0	0	0
14	I like that Product A has a lot of storage. Also, the interface is very easy to use.	...	...	...	...	...	0	0	1	1	0
15	It holds a ton of music.	... Product A	...	...	...	...	0	0	0	0	1
16	It's fun to use	... Product A	...	...	...	...	0	0	0	0	0
17	its cool	... Product A	...	...	...	...	0	0	0	0	0



**Konzepte**

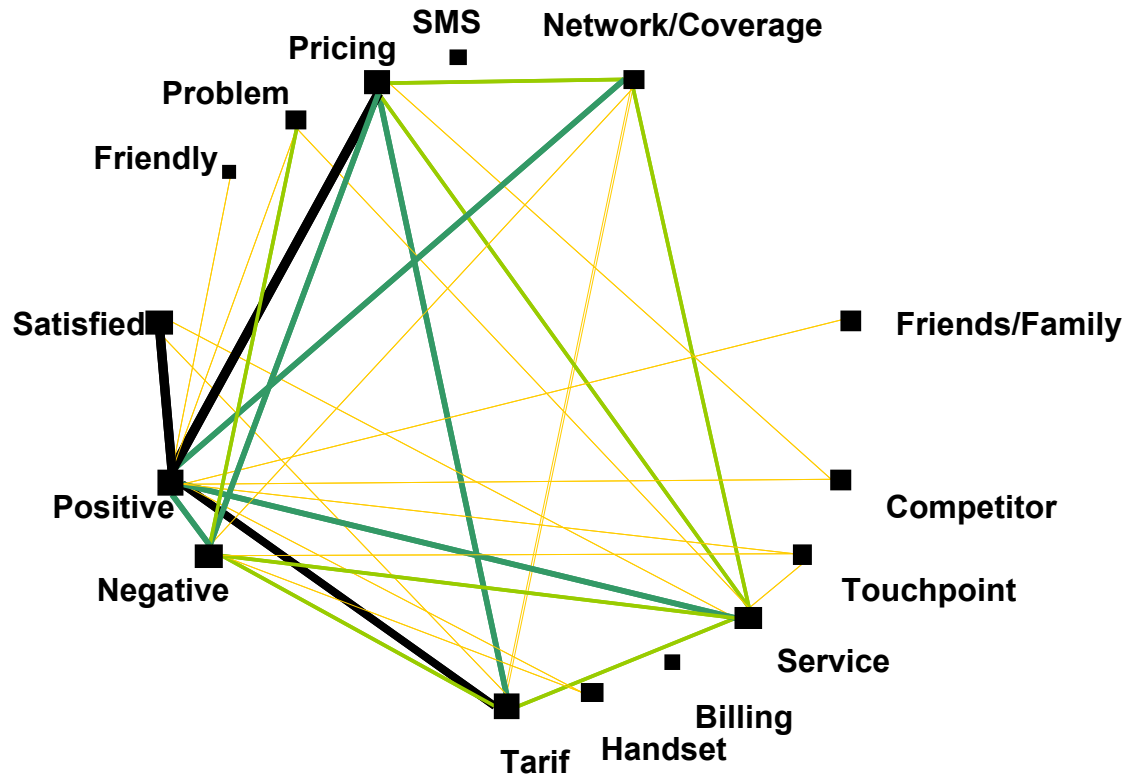
**Klassifikation in die Kategorien „Software“ und „Hardware“**

# Fallstudie: Weiterempfehlung Mobilfunkanbieter

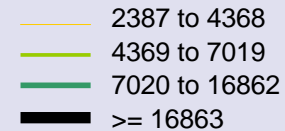
## Ergebnisbeispiel: Netzknotten



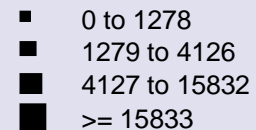
(fiktive Daten)



Absolute Häufigkeit der gemeinsamen Nennungen



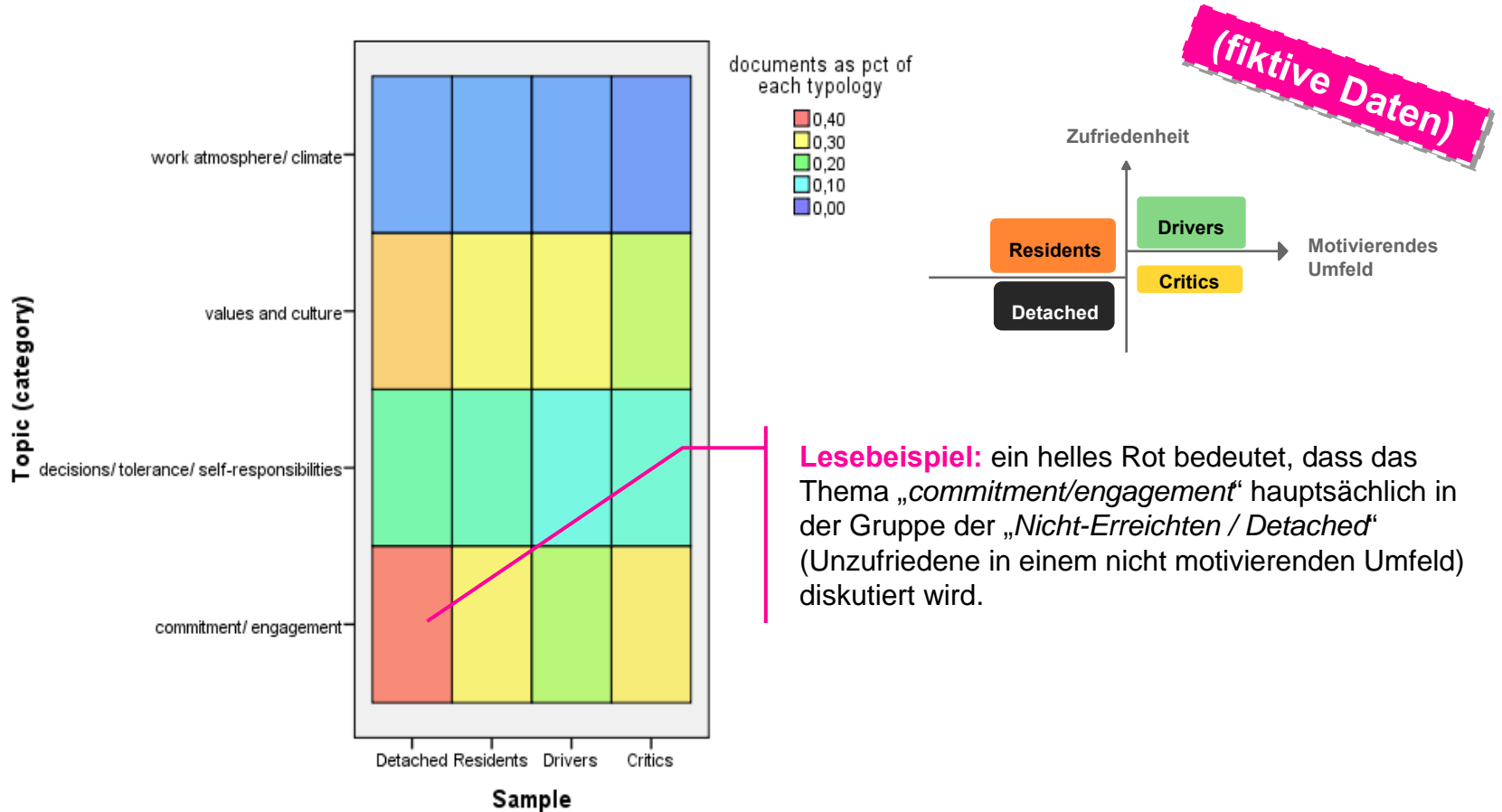
Absolute Häufigkeit der Nennungen



- **Netzknotten** zeigen, welche Begriffe häufig zusammen genannt werden. Dicke Linien deuten auf starke, dünne Linien auf schwache Zusammenhänge hin.

# Fallstudie: Mitarbeiterbefragung

## Ergebnisbeispiel: Heatmaps



- **Heatmaps** bieten eine Möglichkeit, Beziehungen zwischen Nennungen in Untergruppen (hier in den vier TRI\*M-Segmenten) grafisch darzustellen.

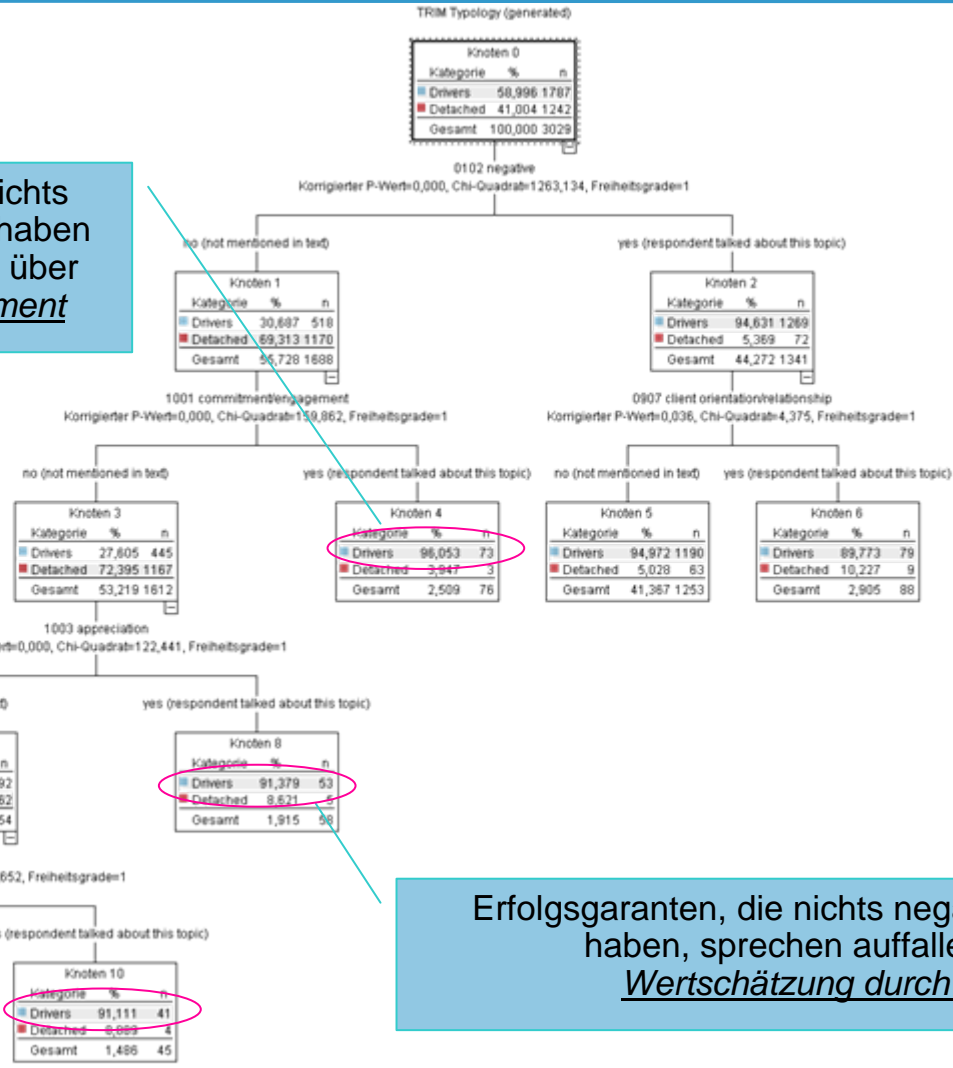
# Fallstudie: Mitarbeiterbefragung

## Ergebnisbeispiel: Treiberanalysen



**(fiktive Daten)**

Erfolgsgaranten, die nichts negatives anzumerken haben sprechen auffallend oft über Commitment/Engagement

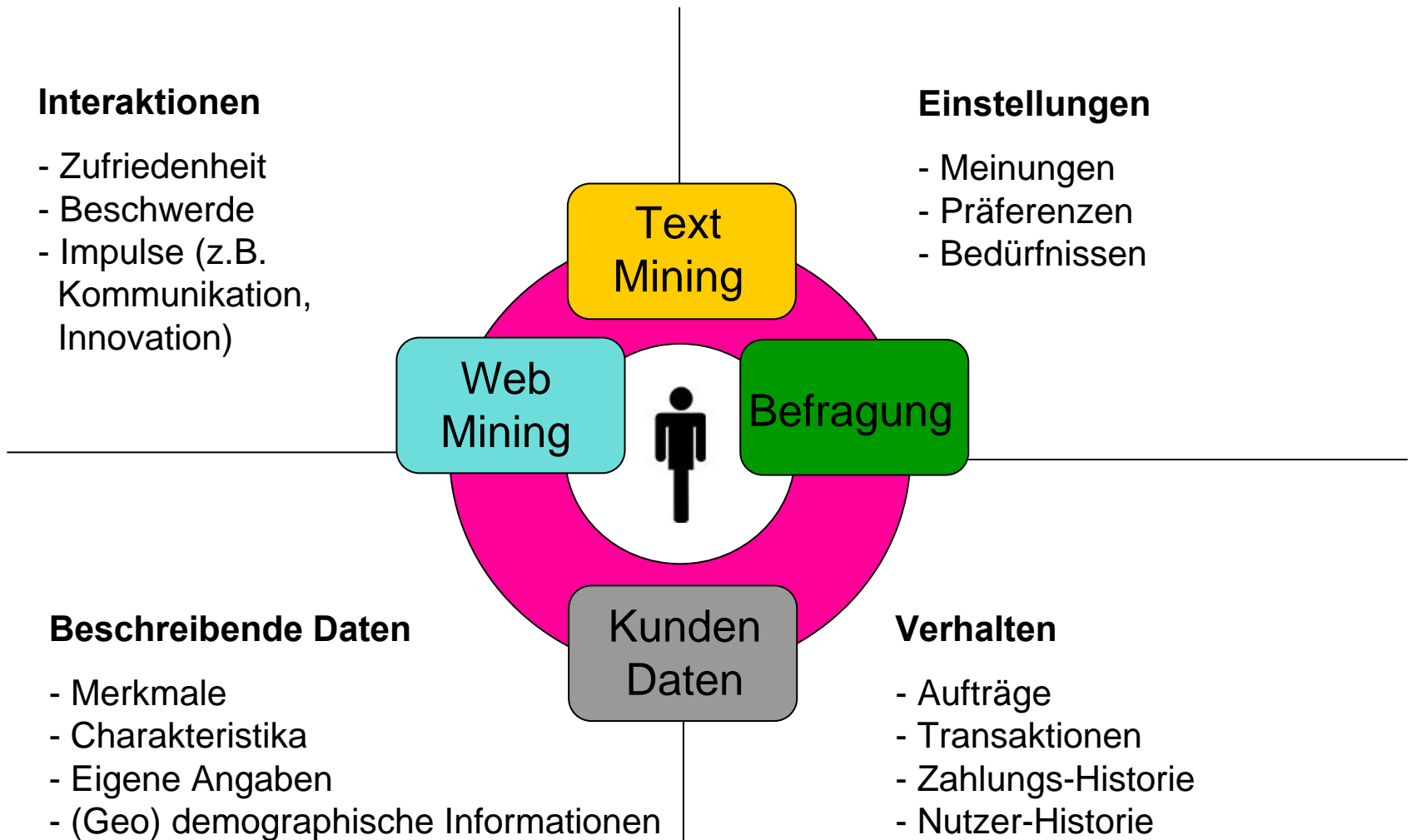


Erfolgsgaranten, die nichts negatives anzumerken haben, sprechen auffallend oft über Wertschätzung durch Kollegen



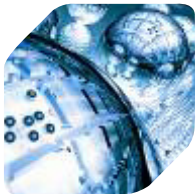
- Text Mining ist ein mächtiges Hilfsmittel, um aus unstrukturierten Daten Informationen und Hypothesen zu extrahieren, zu systematisieren und zu quantifizieren
- Größte Vorteile im Vergleich zur klassischen Verkodung von Verbatims
  - Möglicher Detaillierungsgrad
  - Effizienz
  - Reliabilität und Stabilität auch bei Wiederholungsstudien
  - Monitoring (unmittelbares Aufdecken von neuen Themen und Trends)
- Ergebnisse können vielfältig grafisch aufbereitet werden
- Gewonnene Erkenntnisse können mit Hilfe multivariater Verfahren weiterverarbeitet werden (z.B. Clusteranalysen, MDS, Mappings) – auch in Kombination mit strukturierten Daten
- Die klassische Denkrichtung der Marktforschung ist “top-down” (Suche nach Antworten auf ein spezifisches Problem) – mit Text Mining ist daneben auch Exploration (bottom-up) möglich!
- Generierung geschäftskritischen Wissens, das für das analytische CRM nutzbar wird

# Holistische Sicht auf Kunden Herzstück in einem vorausschauenden Unternehmen

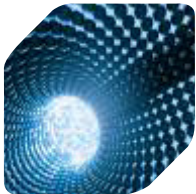




- **Competence Centre** für Data Mining und Data Fusion von TNS Infratest



- Unterstützung aller TNS Marktgesellschaften in den Bereichen **Data Mining**, **Data Fusion** und **Holistic Segmentation** für analytisches CRM und Database Marketing



- **Hoch spezialisierte (Senior) Consultants** und (Senior) Data Analysts aus den Fachgebieten Wirtschaftswissenschaften, Statistik und EDV



- Langjährige **branchenübergreifende und internationale Projekterfahrung** mit komplexen Analysedesigns und Studienkonzepten
- Toolbox mit **state-of-the-art Verfahren** aus den Bereichen der klassischen multivariaten Statistik und des Data Mining
- Enge Kontakte zu Forschungseinrichtungen und Softwareindustrie; fortlaufende Anpassung der Methoden und Tools an den **aktuellen Stand der Forschung**



## ■ **Multivariate Statistik**

- Logistische, kategoriale, lineare Regression, EM-Algorithmus
- Multivariate Adaptive Regression Splines (MARS)
- Ridge Regression, Robust Regression
- Clusteranalyse, Latent Class Analyse

## ■ **Entscheidungsbäume/-regeln, maschinelles Lernen**

- C&RT, C5.0, QUEST, CHAID, Assoziationsregeln
- MART – Multiple Additive Regression Trees, Random Forest
- Nearest Neighbours / Instance based learning EX▪A▪MINE Profiler

## ■ **Künstliche Neuronale Netze**

- Cascade Correlation Learning Architecture, MLP, SOM

## ■ **Hybride Methoden**

- Automatisierte OLAP Navigation und Suche
- Genetische Algorithmen zur Variablenauswahl
- Neuro Fuzzy Algorithmen, interaktive Datenvisualisierung



Holistic  
Customer  
Understanding  
[EX·A·MINE  
Services]

**Dr. Robert Hartl**  
Tel. 089 5600 – 1320  
[robert.hartl@tns-infratest.com](mailto:robert.hartl@tns-infratest.com)

