

# TNS EX·A·MINE™

## TextMiner

Exploration of Relations and  
Trends in Unstructured Data





- In the last years **information overload** has multiplied
- **Unstructured or half-structured** information
- Share of unstructured information: **80 to 90 %**
- How can we capture, analyze, and exploit this information?
- **Text Mining** is a tool to analyze text material / text files
- The **Objective** of Text Mining is:
  - To identify **key concepts and topics**
  - To detect **relations and trends** in the data
- Text Mining is an **automated content analysis**, based on a combination of **linguistic** and **statistical approaches**



## ■ Market research

- Identification of main issues in answers to open-ended questions
- Which issues do the respondents evaluate positively / negatively?
- Which concepts are commonly used in combination?
- Which concepts can be summarized to master categories?

## ■ Customer Relationship Management (CRM)

- Consideration of all possible customer touch points (e-mails, call center contacts, transactions, surveys, ...)
- Which are the primary concerns, needs and problems of customers?
- Which complaints and problem constellations occur in combination ?

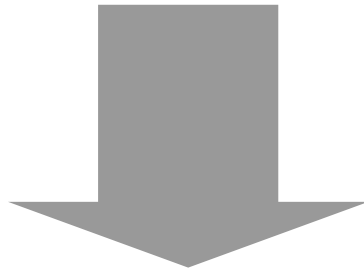
## ■ Analysis of blogs, forums and web-feeds

- Exploration of key subjects emerging on news-feeds, blogs, forums, etc.
- How can these issues be summarised to categories?
- Competitor monitoring and creation of ideas by using blogs

## ■ Fraud detection, medical progress, ...



## Identification of concepts



## Categorization



## Definition of master categories



## Target group analysis

- Concept = word or combination of words
- Analysis based on linguistics
  - Focus on words with meaning (nouns, verbs, adjectives, adverbs)
  - Systematic exclusion of words like articles, preposition, pronouns, etc.
  - Inclusion of root deflection (conjugations/declinations, ...)
- Bundling of concepts with similar content to categories
- Further bundling to more actionable master categories
- Analysis of subgroups using descriptive, bi- and multivariate analyses (classification, association analysis, cluster analysis, driver analysis, ...)



## Algorithms applied to the raw concepts:

- 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*)

# Categorization model

## Classification of data sets / documents



Res...	Q1_What_do_you_like_most_about_this_portable_music_player	... REF1_Pr...	...	...	...	...	volume	mobility	software	...	...	cost	
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	0	...	1	0
8	portability, capacity, sound quality, durability	i...	Other	...	...	...	0	1	0	0	...	0	0
9	Small, great sound, capacity.	...	Product A	...	...	...	0	0	0	0	...	0	0
10	Able to hold all of my songs in one place.	l...	Product A	...	...	...	0	0	0	0	...	0	0
11	It's portable! I can take it anywhere.	...	Product A	...	...	...	0	0	0	0	...	0	0
12	Living in my own little world	...	Product A	...	...	...	0	0	0	0	...	0	0
13	mobility	...	Product A	...	...	...	0	1	0	0	...	0	0
14	I like that Product A has a lot of storage. Also, the interface is very easy to use.	...	Product A	...	...	...	0	0	1	1	...	0	0
15	It holds a ton of music.	...	Product A	...	...	...	0	0	0	0	...	1	0
16	It's fun to use	...	Product A	...	...	...	0	0	0	0	...	0	0
17	its cool	...	Product A	...	...	...	0	0	0	0	...	0	0

**(fictitious data)**

storage, interface

**Concepts**



1, 1

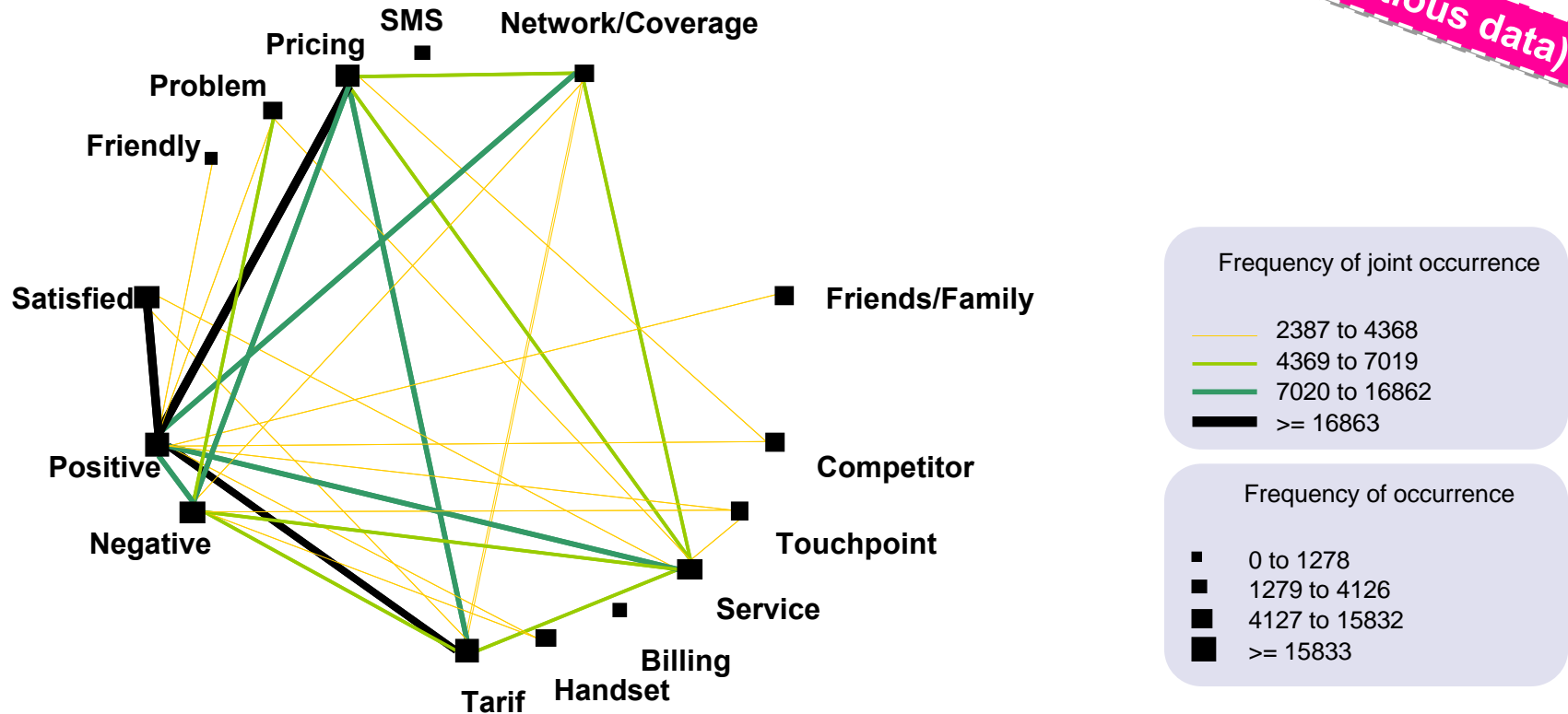
**Assignment to the categories „software“ and „hardware“**

# Case study: Recommendation of mobile phone provider

## Output example: Web graph



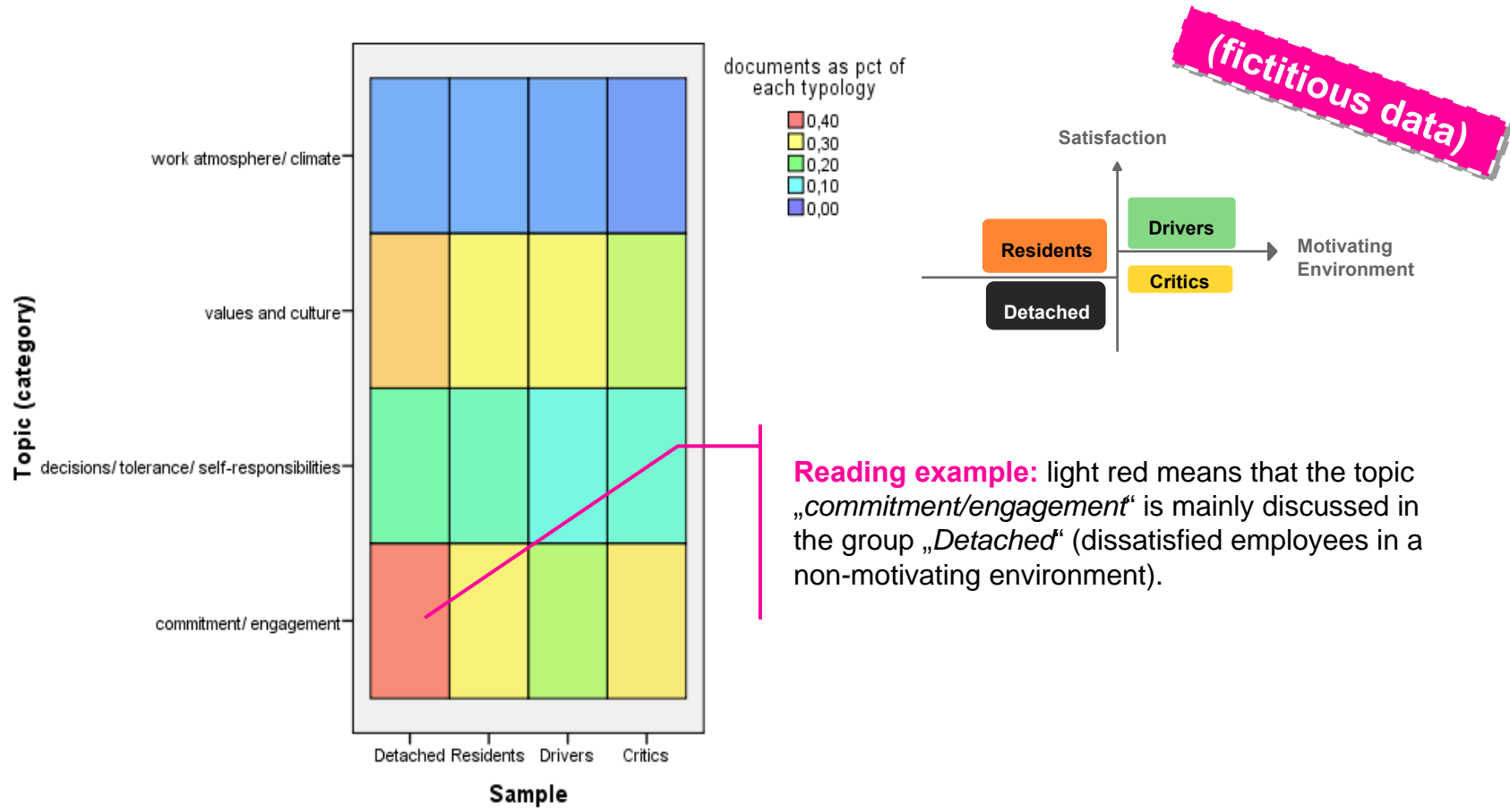
(fictitious data)



■ **Web graphs** show which verbatims frequently occur conjointly. Thick lines represent strong associations, thin lines indicate weak connections.

# Case study: Employee survey

## Output Example: Heatmaps



■ **Heatmaps** provide another way to graphically explore relationships between topics and subgroups (here: the four TRI\*M segments).

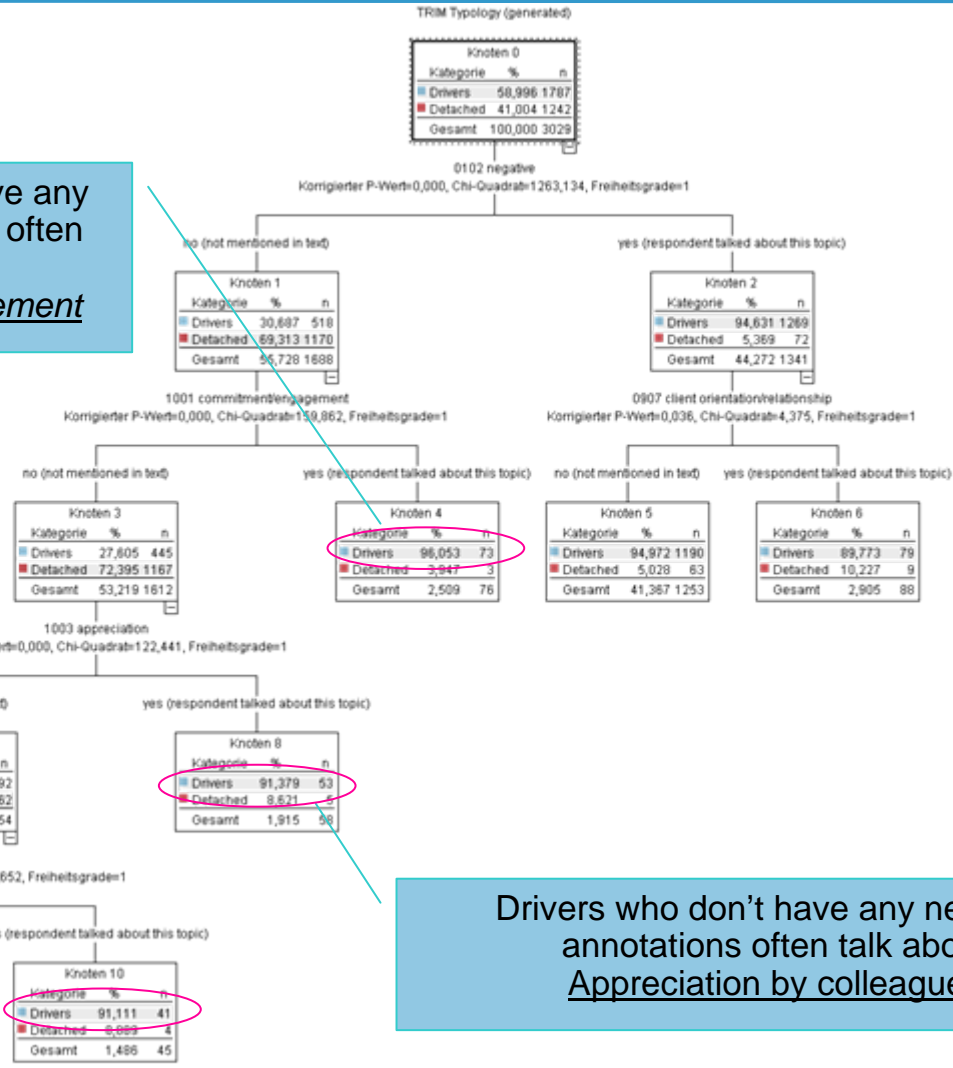
# Case study: Employee survey

## Output example: Driver analysis



**(fictitious data)**

Drivers who don't have any negative annotations often talk about Commitment/Engagement



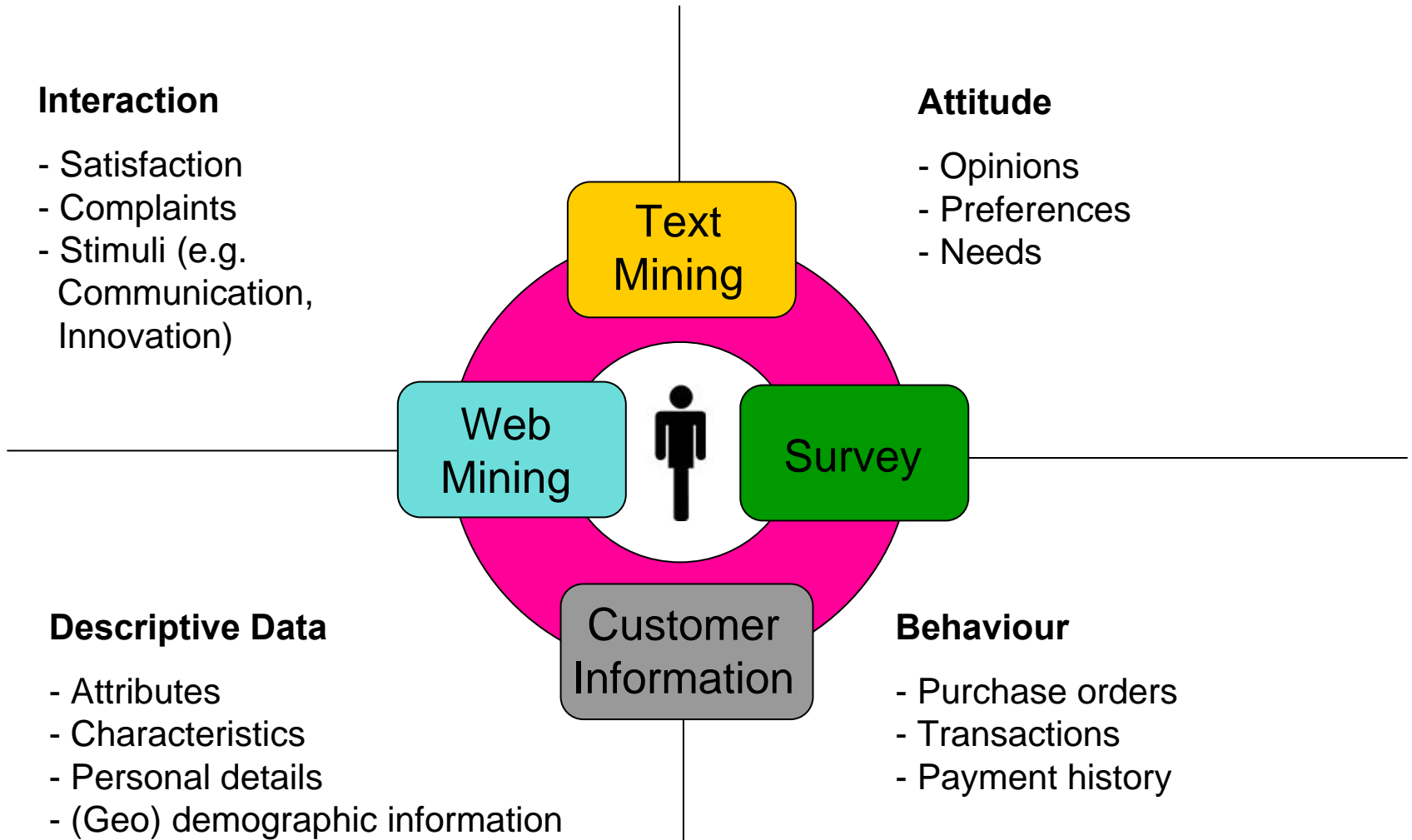
Drivers who don't have any negative annotations often talk about Appreciation by colleagues



- Text Mining is one of the most powerful tools to extract, systemize and quantify new insights and hypotheses from unstructured data
- Advantages compared to conventional coding of verbatims
  - Possible grade of detail
  - Efficiency
  - Reliability and stability in follow up analyses
  - Monitoring (immediate detection of new issues and trends)
- Numerous graphical tools can be used to illustrate the results
- The findings can be further processed by multivariate data analysis approaches (e.g. Cluster Analysis, MDS, Mappings) – even in combination with structured data
- Market researchers are usually thinking top-down and look for topics they are interested in while Text Mining allows for new insights bottom-up
- Generation of business critical knowledge which is also usable in the framework of analytical CRM

# Holistic View on the Customer

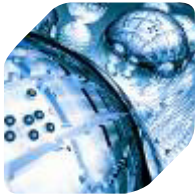
## Core of a forward-looking company



# TNS EX·A·MINE™ at a Glance



- **Competence Centre** for Data Mining and Data Fusion of TNS Infratest



- Support of all TNS Market Sectors in the areas of **Data Mining**, **Data Fusion** and **Holistic Segmentation** for analytical CRM and Database Marketing



- **Highly specialised (Senior) Consultants** and (Senior) Data Analysts whose areas of expertise include Economics, Statistics and IT



- Long **intersectoral and international project experience** with complex analytic designs and study conceptions
- Toolbox with **state-of-the-art methods** in the areas of classic multivariate statistics and Data Mining
- Close relations with research facilities and software industry; ongoing adjustment of methods and tools to the **state-of-the-art**

# The TNS EX·A·MINE™ Algorithms-Toolbox

- **Multivariate statistics**
  - Logistic, Categorical, Linear Regression, EM Algorithm
  - Multivariate Adaptive Regression Splines (MARS)
  - Ridge Regression, Robust Regression
  - Cluster Analysis, Latent Class Analysis
- **Decision Trees / Decision Rules, Automatic Learning**
  - C&RT, C5.0, QUEST, CHAID, Association rules
  - MART – Multiple Additive Regression Trees, Random Forest
  - Nearest Neighbours / Instance based learning EX·A·MINE Profiler
- **Artificial Neural Networks**
  - Cascade Correlation Learning Architecture, MLP, SOM
- **Hybrid Methods**
  - Automatic OLAP Navigation and Search
  - Genetic Algorithms for variable selection
  - Neuro Fuzzy Algorithms, interactive visualisation of data



Holistic  
Customer  
Understanding  
[EX·A·MINE  
Services]

**Dr. Robert Hartl**  
Tel. +49 89 5600 – 1320  
[robert.hartl@tns-infratest.com](mailto:robert.hartl@tns-infratest.com)

