

Q

Business Intelligence



**Data Mining
Techniken und
Anwendungsbereiche**

Reinhard Dechent, Compaq
Michael Wetzel, 

COMPAQ www.compaq.de

Agenda

- ◆ Was ist Data Mining?
- ◆ Methoden und Techniken
- ◆ Durchführung des Data Mining
- ◆ Einsatzbereiche
- ◆ Compaqs Leistungsspektrum

COMPAQ www.compaq.de

Was ist Data Mining?

The process of extracting meaningful information from large databases - information that reveals hidden patterns, trends, and correlations that enables you to make predictions that solve business problems.

Quelle: Thinking Machines

COMPAQ 3 www.compaq.de

Das Ziel



COMPAQ 4 www.compaq.de

OLAP versus Data Mining



OLAP (top down)
Hängt die Kündigung eines Vertrags vom Alter des Kunden ab?



Data Mining (bottom up)
Welche Eigenschaften eines Kunden lassen auf eine bevorstehende Vertragskündigung schließen?

COMPAQ 5 www.compaq.de

Was Data Mining NICHT ermöglicht

- ◆ Antworten auf nicht gestellte oder unrealistische Fragen finden
- ◆ Verzicht auf Datenqualität (GIGO und NINO)
- ◆ Verzicht auf exakte Kenntnis der zugrundeliegenden Daten
- ◆ ... und schon gar nicht ein tiefes Verständnis des Business!

COMPAQ 6 www.compaq.de

Agenda

- ◆ Was ist Data Mining?
- ◆ **Methoden und Techniken**
- ◆ Durchführung des Data Mining
- ◆ Einsatzbereiche
- ◆ Compaqs Leistungsspektrum

COMPACT

www.compaq.de

Data Mining Methoden und Techniken

- ◆ Visualisierende Methoden
 - Häufung (Clustering) und Klassifizierung
 - ...
- ◆ Nicht-visualisierende Methoden
 - Regressionsanalyse
 - Entscheidungsbaum
 - Neuronales Netzwerk
 - Assoziationsanalyse
 - ...

COMPACT

8

www.compaq.de

Beispiel: Häufung und Klassifizierung

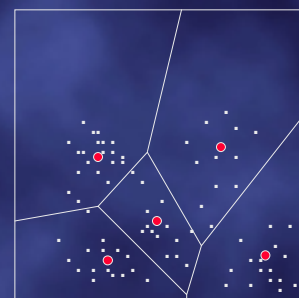


Quelle: Pieter Adriaans, 1996

COMPACT

www.compaq.de

Beispiel: Häufung und Klassifizierung



COMPACT

www.compaq.de

Häufung und Klassifizierung

- ◆ Vorteile
 - ermöglicht Data Mining für alle Arten von Variablen
 - Einfache Anwendbarkeit
- ◆ Schwächen
 - u.U. schwierige Festlegung der richtigen Distanz
 - Interpretation der Cluster

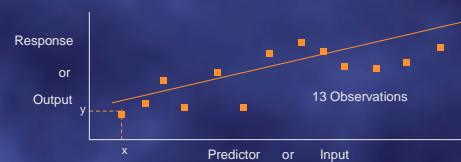
COMPACT

11

www.compaq.de

Die Basis: Lineare Regression

- ◆ Verfahren, mit dem eine abhängige Variable (y) als Funktion des Prädiktors (x) prognostiziert wird
- ◆ Nutzt Standardabweichung und R^2



$$y \equiv \beta_0 + \beta_1 x$$

COMPACT

www.compaq.de

Lineare Regression

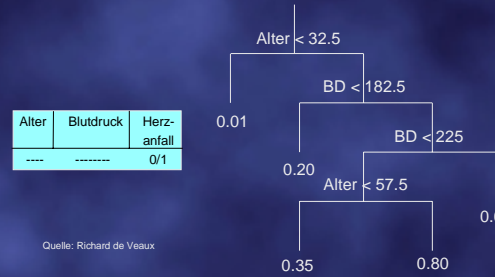
- Vorteile
 - Einfache, nachvollziehbare Darstellung der Ergebnisse
 - messbare Korrelation zwischen tatsächlichem Ergebnis und Vorhersage ($0 \leq R^2 \leq 1$)
- Nachteile
 - Nur mit metrischen Prädiktoren möglich
 - Erfasst nur lineare Abhängigkeiten
 - Es wird vorausgesetzt, dass Messfehler und nicht beobachtete Einflüsse nicht mit den Prädiktoren zusammenhängen

COMPAG

13

www.compaq.de

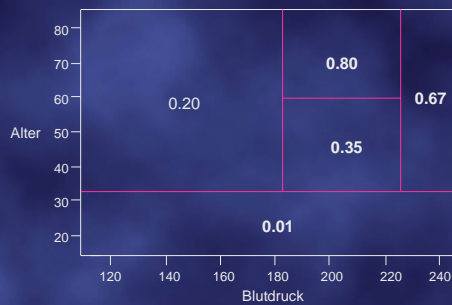
Entscheidungsbaum



COMPAG

www.compaq.de

Beispiel: Entscheidungsbaum



COMPAG

www.compaq.de

Entscheidungsbaum

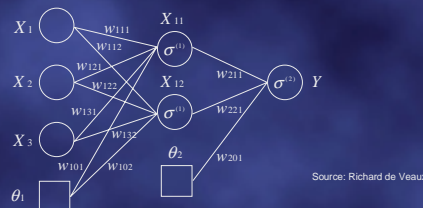
- Vorteile
 - Ergebnisse sind leicht verständlich und nachvollziehbar
 - Anwendbar sowohl auf metrische und diskrete Attribute
 - Verzweigt automatisch nach den Attributen, die die beste Selektion ermöglichen
- Nachteile
 - Erfordert viel Rechenaufwand
 - Gefahr der zu kleinen Segmente (over-fitting)
 - Am besten geeignet für binäre Zielvariablen

COMPAG

16

www.compaq.de

Neuronales Netz



$$y_j = \sigma_j \left(\sum_{k=1}^L w_{2kj} \sigma_k \left(\sum_{j=1}^K w_{1jk} x_j + \theta_j \right) + \theta_k \right)$$

COMPAG

www.compaq.de

Neuronales Netz

- Vorteile
 - Ermöglicht das Auffinden beliebiger Muster
 - Gut geeignet für Vorhersage-Modelle
 - Lernfähigkeit
- Nachteile
 - Ergebnisse sind schwer erklärbar bzw. nicht nachvollziehbar
 - Rechenintensiver Lernprozess
 - Gefahr des Übertrainierens (over-fitting)

COMPAG

18

www.compaq.de

Warenkorbanalyse (Assoziationsregeln)

Suche nach statistisch signifikanten Assoziationsregeln
(der Form if-then)
in den vorhandenen Daten

| A | B | C | D |
|---|---|---|---|
| 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |

If A then C: confidence 50%, support 66%

If A, B then C: confidence 100%, support 33%

Beispiel: MYCIN



www.compaq.de

Warenkorbanalyse: Beispiel

| Relationen | LHS | Support(%) | Confidence(%) | Items |
|------------|-----|------------|---------------|--|
| 34 | 3 | 2.07 | 6.76 | 34.37 (Baby Food ==> Ready Meals & Meat) |
| 55 | 5 | 1.26 | 5.91 | 35.33 (Meat ==> Ready Meals & Fish) |
| 44 | 3 | 1.07 | 6.76 | 36.18 (Ready Meals ==> Meat & Baby Food) |
| 35 | 3 | 1.72 | 6.76 | 34.37 (Baby Food ==> Ready Meals & Fish) |
| 56 | 5 | 1.10 | 5.91 | 61.77 (Ready Meals & Fish ==> Fruit & Veg) |
| 47 | 3 | 1.26 | 6.76 | 72.89 (Ready Meals & Meat ==> Fruit & Veg) |
| 36 | 3 | 1.00 | 6.76 | 28.81 (Fish & Baby Food ==> Ready Meals) |
| 59 | 5 | 1.26 | 5.91 | 61.77 (Ready Meals & Fish ==> Meat) |
| 46 | 3 | 1.03 | 3.26 | 65.86 (Desserts & Meat ==> Fish) |
| 41 | 4 | 1.18 | 16.77 | 100.00 (Meat & Fruit & Veg & Baby Food ==> Fish) |
| 42 | 4 | 1.07 | 16.77 | 59.89 (Meat & Fruit & Veg & Fish ==> Only Food) |
| 43 | 4 | 2.76 | 16.77 | 87.67 (Fish & Baby Food ==> Meat & Fruit & Veg) |
| 44 | 4 | 2.89 | 16.77 | 64.83 (Fruit & Veg & Fish ==> Meat & Baby Food) |
| 45 | 4 | 2.50 | 16.77 | 31.74 (Meat & Baby Food ==> Fruit & Veg & Fish) |
| 48 | 4 | 2.07 | 16.77 | 49.83 (Meat & Fish ==> Fruit & Veg & Baby Food) |
| 47 | 4 | 2.35 | 16.77 | 43.29 (Meat & Fruit & Veg ==> Fish & Baby Food) |



20

www.compaq.de

Warenkorbanalyse

- ◆ Vorteile
 - Einfache, nachvollziehbare Darstellung der Ergebnisse
 - relativ geringer Aufwand in der Datenaufbereitung
- ◆ Nachteile
 - Arbeitet am besten mit wenigen (<100) Variablen
 - Interessante Muster können durch triviale Muster überlagert werden



21

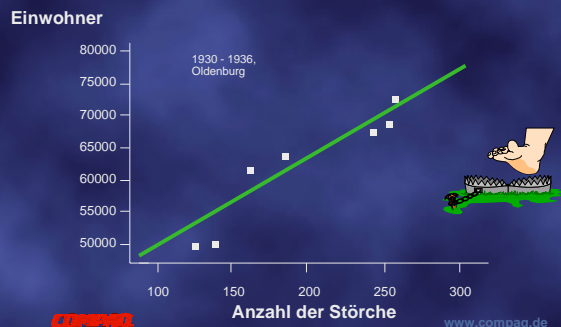
www.compaq.de

- ◆ Was ist Data Mining?
- ◆ Methoden und Techniken
- ◆ **Durchführung des Data Mining**
- ◆ Einsatzbereiche
- ◆ Compaqs Leistungsspektrum



www.compaq.de

Vorsicht ...



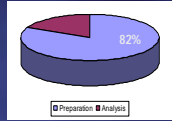
www.compaq.de



www.compaq.de

Der Compaq Data Mining Process

- ◆ Data Mining Pilot
 1. Define the business problem
 2. Load the data
 3. Profile and understand the data
 4. Derive attributes
 5. Define events
 6. Create Data Mining view
 7. Create models
- ◆ Technology Transfer
 8. Deploy models
 9. Measure results



"The assumption that the simple purchase of a data mining package will allow a firm to now "magically" analyze its data is a recipe for failure"

Gartner Group, April 17, 1998



25

www.compaq.de

- ◆ Was ist Data Mining?
- ◆ Methoden und Techniken
- ◆ Durchführung des Data Mining
- ◆ Einsatzbereiche
- ◆ Compaqs Leistungsspektrum



www.compaq.de

Data Mining Verfahren und Einsatz

| Kategorie und Analyseziel | Data Mining Verfahren | Beispiele für praktischen Einsatz |
|--|---|--|
| Klassifikation, Segmentierung, Zuordnung von Individuen in vordefinierte Klassen; „supervised learning“ | Entscheidungsbäume, Memory Based Reasoning, Rule Induction, Neuronale Netze | Creditscoring, Bonitätsbewertung, Kundensegmentierung |
| Zusammenhangs-/Ähnlichkeitsanalysen; Identifikation und Deskription homogener Gruppen; „unsupervised learning“ | Clusteranalyse, Neuronale Netze | Zielgruppenbildung, Lifestyle-Analysen, Warenkorbanalyse, Database Marketing |
| Analyse von Verlaufskurven und Zeitreihen | Neuronale Netze, Fuzzy Logic, Assoziationsanalyse | Verhalten von Bankkunden, Telecom-Nutzern, Konsummuster, Life-Cycle-Analyse |
| Prognoseverfahren | Neuronale Netze, Ähnlichkeits- und Link-Analyse | Zukünftiges Kaufverhalten, Haltbarkeit von Kunden, Prognose der Kaufwahrscheinlichkeit |
| Beschreibung, Charakterisierung von Kunden, Produkten, usw. | Visualisierung, deskriptive und exploratorische Analysen | Analyse von Zielgruppen, Untersuchung typischer Merkmale, Taxonomien, Kundenprofitabilität |



www.compaq.de
www.systematika.com

Data Mining bei Finanzdienstleistern

- ◆ Marktbearbeitung, Marketing
- ◆ Cross-Selling durch Mailings und Berater
- ◆ Betrugserkennung
- ◆ Credit Scoring
- ◆ Frühwarnsysteme
- ◆ Portfolioanalyse



www.compaq.de
www.systematika.com

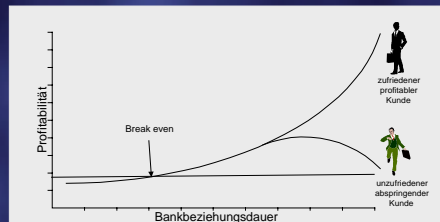
Banken: Retention Marketing

- ◆ Absprunggefährdete Kunden halten
- ◆ Dauer der Kundenbeziehung verlängern
- ◆ Kundenbeziehung intensivieren
- ◆ Deckungsbeiträge verbessern
- ◆ Produktpolitik optimieren



www.compaq.de
www.systematika.com

Banken: Retention Marketing



Kundenloyalität und Profitabilität sind korreliert
Neukunden gewinnen ist 10x teurer als Altkunden halten
Absprunggefährdung ist Monate vorher erkennbar



www.compaq.de
www.systematika.com

Banken: Retention Marketing Projekte

Projekttablauf

- Bankstrategisches Konzept
- Marktforschung: Gründe für Weggang
- Aufbau der Datenbasis, Data Cleaning
- Datenexploration, Hypothesenprüfung
- Entwicklung eines Betreuungskonzepts für die Kunden
- Modell-Entwicklung / Evaluation
- Durchführung von Retention-Marketing-Massnahmen
- Pilotbetrieb, Ergebniskontrolle
- Produktiver Einsatz, Controlling



www.compaq.de
www.systematika.com

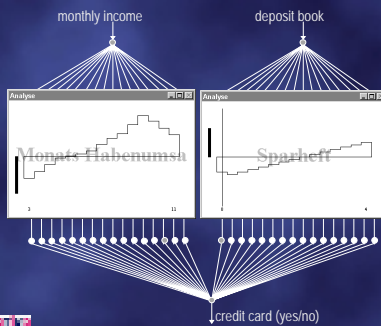
Banken: Retention Marketing Kosten/Ertrag

| Ertrag | | DM |
|--|-------------------------------|------------------|
| 20% Absprungkandidaten können gehalten werden à Marge 300 DM bei 100.000 Kunden und Abgangsrate 5% | 20% von (5000*300 DM) = 3 Mio | 3.000.000 |
| Kosten | | |
| Kosten Datenbelieferung und Berechnung | Pro Jahr | -500.000 |
| Externe Kosten Retention-Projekt | Pro Jahr | -500.000 |
| Retention-Massnahmen für 20.000 Kunden à 10 DM | Pro Jahr | -200.000 |
| Gewinn | Circa | 2.000.000 |



www.compaq.de
www.systematika.com

Banken: Cross-Selling/Direct Mailings



www.compaq.de
www.systematika.com

Banken: Cross-Selling/Neuronale Netze

Discriminant analysis:

| samples | no credit card | credit card | total |
|---------|----------------|-------------|-------|
| 2000 | 83% | 43% | 63% |
| 8000 | 86% | 42% | 64% |

Neural Network:

| samples | no credit card | credit card | total |
|---------|----------------|-------------|-------|
| 2000 | 73% | 75% | 74% |

Ergebnis:

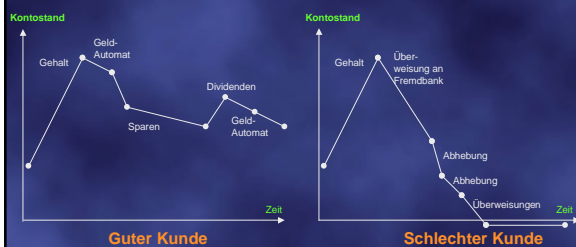
Neuronale Netze haben eine hohe Vorhersagekraft
Neuronale Netze ergänzen statistische Analysen



credit card (yes/no)

www.compaq.de
www.systematika.com

Banken: Prognose des Kundenverhaltens



credit card (yes/no)

www.compaq.de
www.systematika.com

Banken: Database Marketing

Database Marketing ist in Banken sehr wichtig geworden

DBM ist ein kontinuierlicher Optimierungsprozess

Modernste Analysemethoden sind nötig

Life Time Value-Modelle müssen entwickelt werden

Data Mining muss als Business-Prozess verankert sein

Schrittweise sichere Einführung vermindert Risiken



credit card (yes/no)

www.compaq.de
www.systematika.com

Data Mining bei Fluggesellschaften

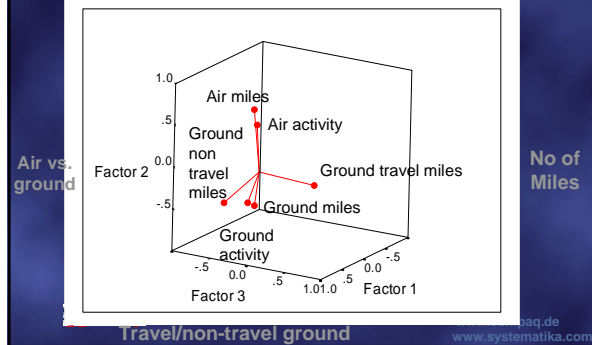
- ◆ Kundenanalyse
- ◆ Produktanalyse
- ◆ Marktanalyse
- ◆ Trends



www.compaq.de
www.systematika.com

Fluggesellschaft: Kundensegmentierung

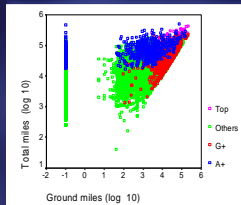
Component Plot



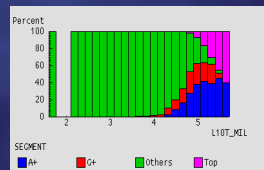
Travel/non-travel ground

www.compaq.de
www.systematika.com

Fluggesellschaft: Kundensegmentierung



| Segments | Top | G+ | A+ | Others |
|-------------|-----|-----|-----|--------|
| n | 100 | 680 | 604 | 11644 |
| % air miles | 70 | 51 | 95 | 96 |
| # air | 95 | 17 | 55 | 7 |
| # ground | 41 | 23 | 5 | 0.7 |



www.compaq.de
www.systematika.com

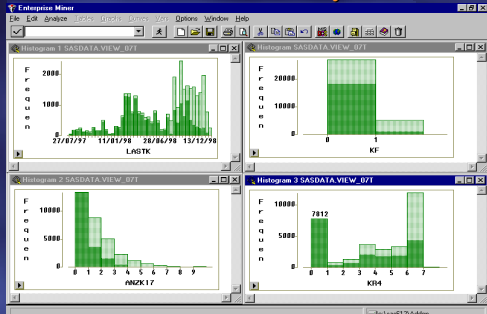
Data Mining im Versandhandel

- ◆ Marktvolumen- und Potential-Analysen
- ◆ Kundenstrukturanalysen, Segmentierung
- ◆ Neukundengewinnung
- ◆ Produktanalysen
- ◆ Cross-Selling
- ◆ Bonitätsbewertung
- ◆ Warenkorbanalysen
- ◆ Absatzprognosen, Trends



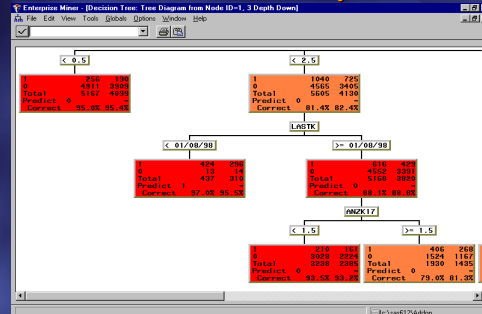
www.compaq.de
www.systematika.com

Versandhandel: Käuferanalyse



www.compaq.de
www.systematika.com

Versandhandel: Käuferanalyse



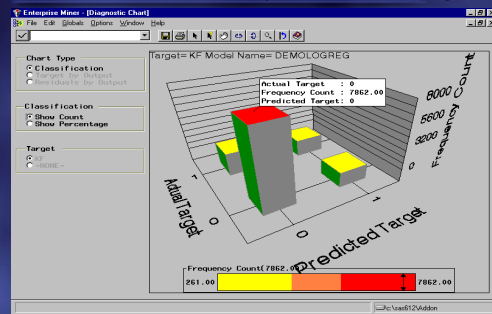
www.compaq.de
www.systematika.com

Versandhandel: Käuferanalyse

| Fit Statistic | Label | Training | Validation | Test |
|---------------|--|---------------|--------------|--------------|
| _ASE | Train Average Squared Error | 0.1113485425 | 0.1306720555 | 0.109100002 |
| _AVERE | Train Average Error Function | 0.3647941130 | 0.3552409004 | 0.3579005887 |
| _DFE | Train Degrees of Freedom for Error | 12720 | | |
| _DFM | Train Model Degrees of Freedom | 18 | | |
| _DFV | Train Total Degrees of Freedom | 12738 | | |
| _DVF | Train Devior for ASE | 25440 | 19404 | 19079 |
| _JFPE | Train Error Function | 8200.3749705 | 6893.0961775 | 7115.9435086 |
| _FPE | Train Final Prediction Error | 0.1116662495 | | |
| _MAX | Train Maximum Absolute Error | 0.8930257471 | 0.897870884 | 0.8973566976 |
| _MSE | Train Mean Squared Error | 0.1113485425 | 0.1306720555 | 0.109100002 |
| _NOBS | Train Sum of Frequencies | 12720 | 8702 | 8659 |
| _NOV | Train Number of Estimate Weights | 18 | | |
| _PASE | Train Root Average Sum of Squares | 0.3336902066 | 0.3327599295 | 0.3303166061 |
| _RFFPE | Train Root Final Prediction Error | 0.3341816305 | | |
| _RMSE | Train Root Mean Squared Error | 0.3332925421 | 0.3207592595 | 0.3303166061 |
| _SBC | Train Schwarz's Bayesian Criterion | 3450.4617565 | | |
| _SSE | Train Sum of Squared Errors | 2632.70629121 | 2071.804084 | 2169.8698956 |
| _SUMM | Train Sum of Case Weights Times Fits | 25440 | 19404 | 19079 |
| _W | Model Classification Rate | 0.1545597484 | 0.1446003589 | 0.1501127050 |
| _WCF | Train Frequency of Classified Cases | 12720 | 8702 | 8659 |
| _WTF | Train Frequency of Misclassified Cases | 0 | 0 | 0 |

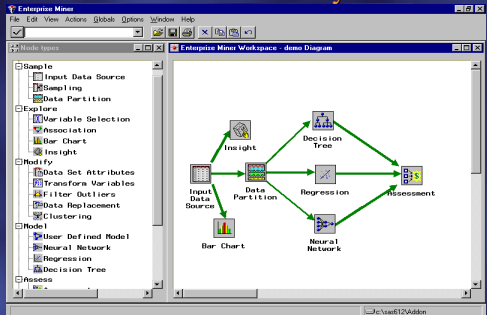
www.compaq.de
www.systematika.com

Versandhandel: Analyse der Genauigkeit



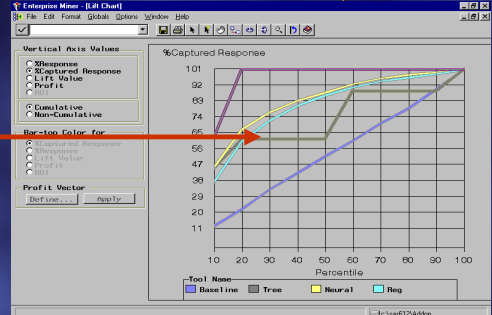
www.compaq.de
www.systematika.com

Versandhandel: Käuferanalyse



www.compaq.de
www.systematika.com

Versandhandel: Käuferanalyse



www.compaq.de
www.systematika.com

Wie wird Data Mining zum Erfolg?

- ◆ Analyse-Team mit hoher Business-Intelligenz und statistischem Verständnis aufbauen
- ◆ Heuristische Analyse in geschäftskritischen Bereichen
- ◆ Data Mining und Analysen popularisieren
- ◆ Analysen mit hohem Return on Investment definieren
- ◆ Geschäftsentscheidungen empirisch absichern
- ◆ Analysen automatisieren, Workflows definieren

www.compaq.de
www.systematika.com

- ◆ Was ist Data Mining?
- ◆ Methoden und Techniken
- ◆ Durchführung des Data Mining
- ◆ Einsatzbereiche
- ◆ Compaqs Leistungsspektrum

COMPAG

www.compaq.de

Data Mining Lösungen Entscheidungen für die Zukunft



Hochleistungs- Plattformen

flexibel, skalierbar, ökonomisch

ProLiant, AlphaServer, Himalaya
Windows NT, Tru64 UNIX
OpenVMS, NonStop Kernel
StorageWorks

Strategische Partnerschaften

mit Software- & Lösungspartnern



Gesamtlösungs- Kompetenz

durch langjährige Erfahrungen

Compaq Services
gemeinsam mit Partnern

COMPAQ

49

www.compaq.de

So wird Data Mining für Sie zum Gewinn

- ♦ Mit der flexibelsten & leistungsfähigsten Plattform
... gebaut von Compaq
- ♦ Mit den besten Werkzeugen für Ihre Anforderungen
... integriert durch Compaq
- ♦ Mit einem strukturierten & pragmatischen Vorgehen
... bereitgestellt von Compaq

Weitere Infos, Success Stories, Demos und Kontakt zu den Experten:



COMPAQ

www.compaq.com/activeanswers

business.intelligence@compaq.com

http://www.compaq.de/loesung/