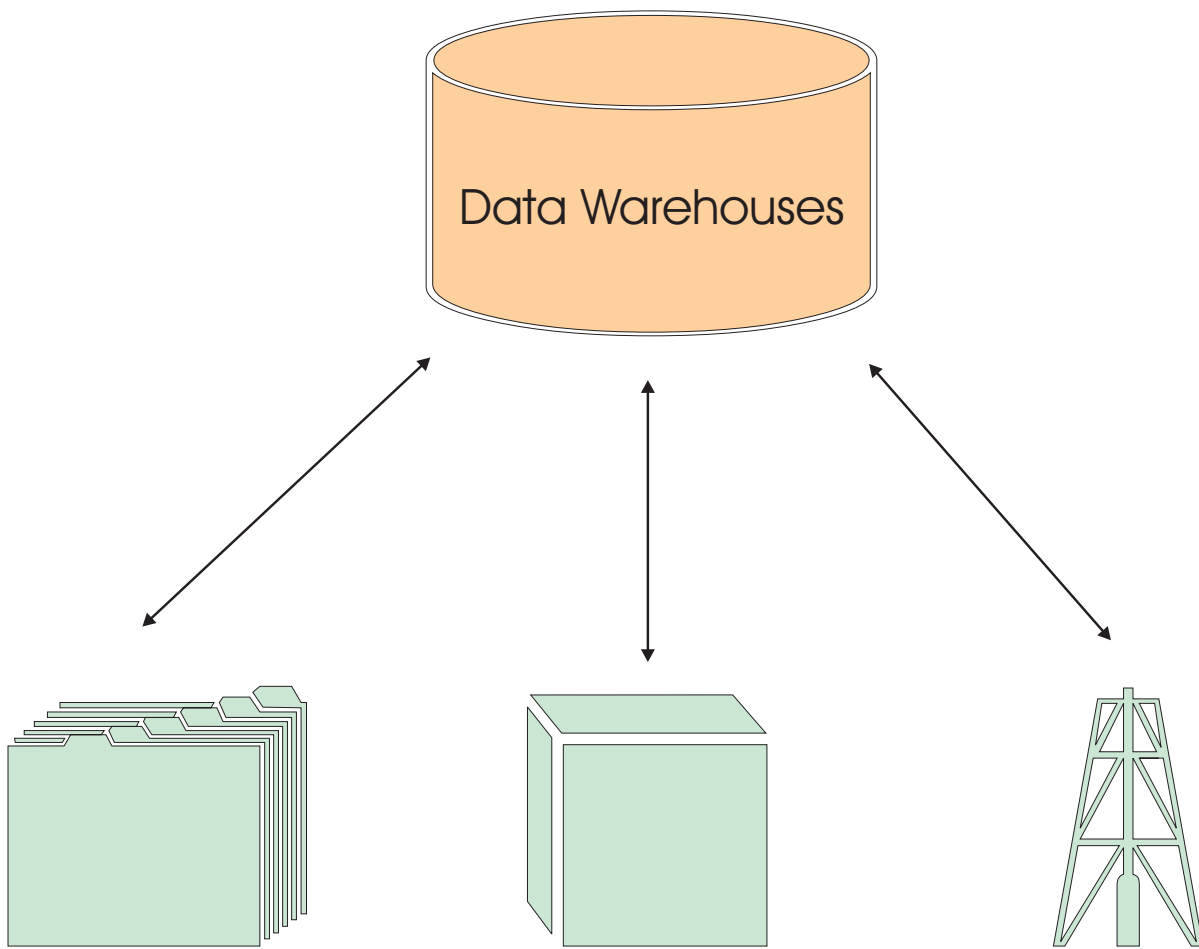

Data Warehousing und Data Mining

Eine Einführung in entscheidungsunterstützende Systeme

Interaktive Folien zu Kapitel 6
Data Mining - ein Überblick

6.1 Data Mining





Abfrage- und Berichtssprachen wie **SQL** und QBE sind standardisiert und mächtig, aber für den gelegentlichen Benutzer zu schwierig

OLAP-Werkzeuge erlauben auch dem gelegentlichen Benutzer flexible multi-dimensionale Abfragen

Data Mining-Werkzeuge lassen den erfahrenen Benutzer in Massendaten mit komplexen Methoden nach verborgenem Wissen "schürfen"

Einordnung





Nutzwertanalyse am Beispiel von AHP

- ✓ Kioskstandort  , Personalauswahl 

Was Wenn-Analyse

- ✓ Erfolgsrechnung 
- ✓ Anzeigenplanung , Produktionsplanung 

Regelbasierte Systeme

- ✓ Spesen , Betriebskredit 
- ✓ Regelverkettung  

Data Warehouses

- ✓ Anlageberatung  
- ✓ Lieferfrist, Handel, Verkauf  

⇒ **Data Mining** - Ein Überblick

- ⇒ Zeitschriften , Bank 

Regelinduktion

- Spesen , Bonitätsklassifikation  

Neuronale Netze

- Bonitätsklassifikation , Bonitätsvorhersage 
- EindimPerzeptron  , ZweidimPerzeptron  
- MehrklassPerzeptron  
- MehrstufPerzeptron  

Einblick in das Data Mining

Data Warehouses

- ✓ Begriff
- ✓ Endbenutzerzugriff
- ✓ Modellierung
- ✓ Entwicklung und Betrieb

Einblick in das Data Mining

⇒ Begriff, Anwendung und Entwicklung	<u>5</u>
⇒ ① Konventionelle Abfragewerkzeuge	<u>15</u>
⇒ ② Konventionelle statistische Analyse	<u>19</u>
⇒ ③ Klassifikation mit Entscheidungsbäumen	<u>21</u>
⇒ ④ Neuronale Netze	<u>22</u>
⇒ ⑤ Visualisierung	<u>26</u>
⇒ Methoden- und Werkzeugauswahl	<u>41</u>
⇒ BANK 🖱	<u>47</u>
⇒ Data Mining und Tabellenkalkulation?	<u>50</u>

Data Mining

to mine for heisst schürfen nach



Data Mining :=

- ✓ nichttriviales
- ✓ automatisches Schürfen
- ✓ nach Wissen
- ✓ in Massendaten



›Data Warehouses als Datenlieferanten

- 1 Synonym: Datenmustererkennung
- 2 nichttrivial: mit Methoden aus ›KI und Statistik
(nicht nur mit ›SQL, ›OLAP und ›Berichtsgeneratoren)
- 3 Massendaten: z.B. Daten über Prospekt-Empfänger

6.2 Anwendungen

<i>Anwendungsbeispiel</i>	<i>Funktionsbereich</i>
Marktkorbanalyse	Absatz
Verkaufsprognose	Absatz, Produktion
Beurteilung der Werbewirksamkeit	Absatz
Antwortrate für Direct Mailing	Absatz
Kreditwürdigkeitsbeurteilung →	Absatz
Entdeckung von Kreditkartenbetrug	Absatz
Kundenzufriedenheit	Absatz
Analyse der Zahlungsgewohnheiten	Rechnungswesen
Mitarbeitererevaluation	Personal
Mitarbeiterzufriedenheit	Personal



Anwendungsklassen

Anwendungsphasen

Anwendung spezifizieren →

› Bonitätsbeurteilung



Anwendungsklasse bestimmen →

› Klassifikation



Methodenklassen vorschlagen →

› Neuronale Netze und › Regelinduktion



Methodenklasse wählen

› Regelinduktion



Werkzeug wählen

NeuralWorks Predict

6.3 Anwendungsklassen

Klasse	Aufgabe	Anwendung	Methoden
Klassifikation	Individuen bekannten Klassen zuordnen	›OCR, Bonitätsbeurteilung	›Regelinduktion, ›Neuronale Netze
Vorhersage	Kontinuierliche zukünftige Werte aus unabhängigen Variablen berechnen	Bonitätsbeurteilung	›Neuronale Netze, ›Regression
Clustering	Gruppen aufgrund von Ähnlichkeiten zwischen Individuen identifizieren	Werbeadressaten einteilen	›Neuronale Netze, konventionelle ›Clusteranalyse
Assoziation	Abhängigkeiten entdecken und quantifizieren	›Marktkorb-analyse ¹	statistische Zusammenhangsanalyse
Text Mining	Textmuster suchen	Information Retrieval	Suchalgorithmen



Bsp. ›Bonitätsbeurteilung

¹ Bsp. 80% aller Kunden, die Bier kaufen, haben auch Chips im Einkaufskorb.

Methodenklassen

Die Information verdoppelt sich alle 20 Jahre



Reduktion (Verdichtung) grosser Datenmengen

Konventionelle Statistik

- › Entscheidungs bäume (Regelinduktion)
- › Neuronale Netze
- › Bayes-Netze
- › Assoziationsanalyse
- › Visualisierung

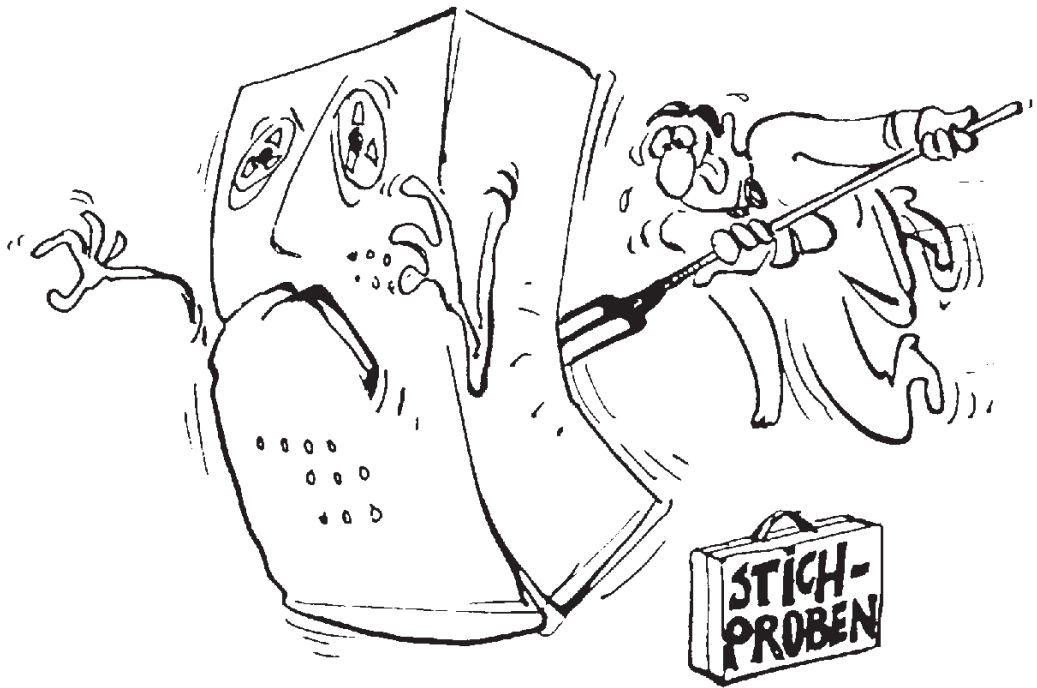
...



⇒ am Fallbeispiel *ZEITSCHRIFTEN*

⇒ allgemein

Fussangel des Data Mining



Die Daten solange quälen, bis sie das hergeben, was man von ihnen erwartet ...

Voraussetzungen des Data Mining

Data Warehouses



Daten

- relevant
- genügend
- zuverlässig



Data Mining



Hypothesen über ...

- wichtige Attribute
- Beziehungen



Betriebliches Fachwissen

Data Mining ist meist datengetrieben

① Datengetriebene Datenanalyse

Synonyme

- entdeckende -
- explorative -

Ausgangspunkt

Daten, deren Muster beschrieben und verallgemeinert werden sollen

Verfahren

- ›Deskriptive Statistik
- ›Visualisierung
- ›Neuronale Netze

② Modellgetriebene Datenanalyse

Ausgangspunkt

Hypothese, die aus einem Modell der Wirklichkeit abgeleitet wurde

Verfahren

- ›Inferenzstatistik
- ›SQL

6.4 Daten als Ausgangspunkt

Begriff	Synonyme	Verwandte Begriffe
<i>Datengesamtheit</i>	Lernmenge	▸ Grundgesamtheit ▸ Stichprobe Testmenge Datenbank
<i>Datenelement</i>	Satz Zeile Tupel Individuum Beispiel	
<i>Variable</i>	Attribut Merkmal Feld Spalte	Fakt Indikator Dimension
<i>Unabhängige Variable</i>	exogene V.	Prädiktor Klassifikator
<i>Abhängige Variable</i>	endogene V. Kriterium	vorhergesagte V. klassifizierte V.

Entwicklungsphasen

① Problem spezifizieren

- Abhängige Variable bestimmen
- Unabhängige Variablen bestimmen

② Daten sammeln und aufbereiten

- Vollerhebung oder Teilerhebung durchführen
- Datengesamtheit ev. in eine Lern- und Testmenge aufteilen
- Variablen messen und transformieren

③ Daten explorieren

- Ausreisser analysieren
- Variablen in Beziehung zueinander setzen (Bsp. ›Korrelation)
- Daten visualisieren (Bsp. Streudiagramm)
- Verteilungsmasse berechnen (Bsp. Varianz)
- Hypothesen formulieren

④ Hauptmethoden anwenden

- Methoden wählen
- Hypothesen testen
- Zusammenhänge identifizieren und quantifizieren
- Kausalität von blosser ›Korrelation unterscheiden
- Einfluss intervenierender Variablen identifizieren

⑤ Ergebnisse validieren, präsentieren und anwenden

- Unsicherheit der Ergebnisse quantifizieren (Bsp. Signifikanztest)
- Ergebnisse visualisieren →
- Ergebnisse auf neue Stichproben anwenden

① Data Mining im weiteren Sinne

› Abfragesprachen und › Berichtsgeneratoren¹

- ☹ beantworten einfache Fragen *abschliessend* und *entdecken Hypothesen*, die später mit Data Mining-Werkzeugen getestet werden können
- ☹ sind kaum automatisierbar
- ☹ berechnen Masse wie Häufigkeiten, Durchschnitte, Standardabweichungen

+

Tabellenkalkulationswerkzeuge

- ☹ nur für geringe Datenmengen
- ☹ Funktionsumfang beschränkt

↓

80% des interessierenden Wissens lässt sich mit konventionellen Werkzeugen gewinnen, die übrigen 20% mit Data Mining-Werkzeugen i.e.S

¹ Bsp. SQL, QBE, *Crystal Reports*

6.5 📌 ZEITSCHRIFTEN - Data Mining i.e.S.

Data Warehouses (Teil I des Fallbeispiels)

Ein Verlag publiziert die Zeitschriften Auto, Wohnen, Sport, Musik und Comics. Ein Data Warehouse soll Fragen der folgenden Art beantworten: “Welches Profil zeigen die Leser der Zeitschrift Wohnen?” oder “Welche Zusammenhänge bestehen zwischen Lesern der Zeitschriften Sport und Auto?”. Data Mining Tools sollen auf dem Data Warehouse aufsetzen und Marketing-Aktionen wie ›Direct Mailing oder Anzeigenkampagnen unterstützen. Das Management stellt sich vor, dass die Analysen zum Beispiel die folgenden Aussagen ermöglichen sollen:

- *“Hypothekarverschuldete Leser sind mit einer Wahrscheinlichkeit von 70% auch Abonnenten der Zeitschrift Wohnen.”*
- *“Ein Abonnent der Zeitschrift Sport mit einem Alter zwischen 20 und 30 subskribiert mit einer Wahrscheinlichkeit von 40% zwischen 30 und 40 auch die Zeitschrift Auto”.*

Data Mining im weiteren Sinn ✓

Data Mining im engeren Sinn

- ⇒ ② Konventionelle statistische Analyse
- ⇒ ③ Induktion von Entscheidungsbäumen
- ⇒ ④ Neuronale Netze
- ⇒ ⑤ Visualisierung

📌 Naive Vorhersage

Frage

Abonniert ein Konsument eine bestimmte Zeitschrift?

Naive Vorhersage

Zeitschrift	Kaufhäufigkeit	Vorhersage
“Auto”	40%	⇒ kein Abonnement
“Wohnen”	70%	⇒ Abonnement
...

Schlussregel

Relative Häufigkeit in der Stichprobe $> 0.5 \Rightarrow$
Merkmal liegt auch in der Grundgesamtheit vor



Naive Vorhersagen untersuchen *keine Faktoren*, welche die vorherzusagende Grösse beeinflussen könnten. Sie berücksichtigen nur die leicht verfügbare *Vorinformation* (**a priori**-Information)

Naive Vorhersagen schätzen **a priori**-Wahrscheinlichkeiten

Eine methodische Vorhersage schätzt hingegen *bedingte* Wahrscheinlichkeiten unter Berücksichtigung ihrer Faktoren →

📌 Methodische Vorhersage

Frage

Abonniert ein Konsument eine bestimmte Zeitschrift?

Methodische Vorhersage

Zeitschrift	Wohneigentum	Autoeigentum	...	Vorhergesagte Abonnements-wahrscheinlichkeit
"Auto"	ja	ja	...	65%
"Wohnen"	ja	nein	...	80%
...

Viele Data Mining - Methoden (zum Beispiel Regressions- oder Entscheidungsbaumverfahren) schätzen die **a posteriori**-Wahrscheinlichkeit einer Vorhersagegrösse unter der Annahme, dass bestimmte **Faktoren** diese Grösse beeinflussen

Data Mining-Verfahren müssen nachweisen, dass ihre Validität (Vorhersagegüte) besser ist als jene naiver Prognosen

② Klassifikation und Clustering in der Statistik

a) Clustering

Gruppen von Stichprobenelementen identifizieren, ohne dass die Gruppenzugehörigkeiten zum voraus bekannt sind

Hierarchische -

für kleinere Stichproben, Clusterzahl nicht erforderlich
Variablen nominal-, ordinal- oder intervallskaliert
flexibler als K-Means-Clusteranalyse

K-Means-

für grössere Stichproben, Clusterzahl erforderlich
Variablen intervallskaliert, effizienter als hierarchische -

b) Klassifikation

Variablen identifizieren, die *bekannte* Gruppenzugehörigkeiten erklären und auf künftige Klassifikationsaufgaben anwenden

▸ Diskriminanzanalyse

unabhängige Variablen multivariat normal verteilt, abhängige Variable nominal skaliert, ausschliessliche Gruppenzugehörigkeiten

Logistische Regression

abhängige Variable dichotom
unabhängige Variablen nominal oder intervallskaliert
weniger einschränkende Annahmen als Diskriminanzanalyse

Lineare ▸ Regression

abhängige Variable stetig

③ Entscheidungsbaum

Forschungsfrage

Welche Bevölkerungsgruppen lesen “Auto und Freizeit”?

Methoden

a) *Einrückungsliste*

Wer liest “Auto und Freizeit” ?

Alter ≤ 45

Einkommen $> 70'000.- \Rightarrow 0\%$

Einkommen $\leq 70'000.-$

Alter $\geq 32 \Rightarrow 52\%$

Alter $< 32 \Rightarrow 95\%$

Alter > 45

Alter $> 50 \Rightarrow 0\%$

Alter $\leq 50 \Rightarrow 8\%$

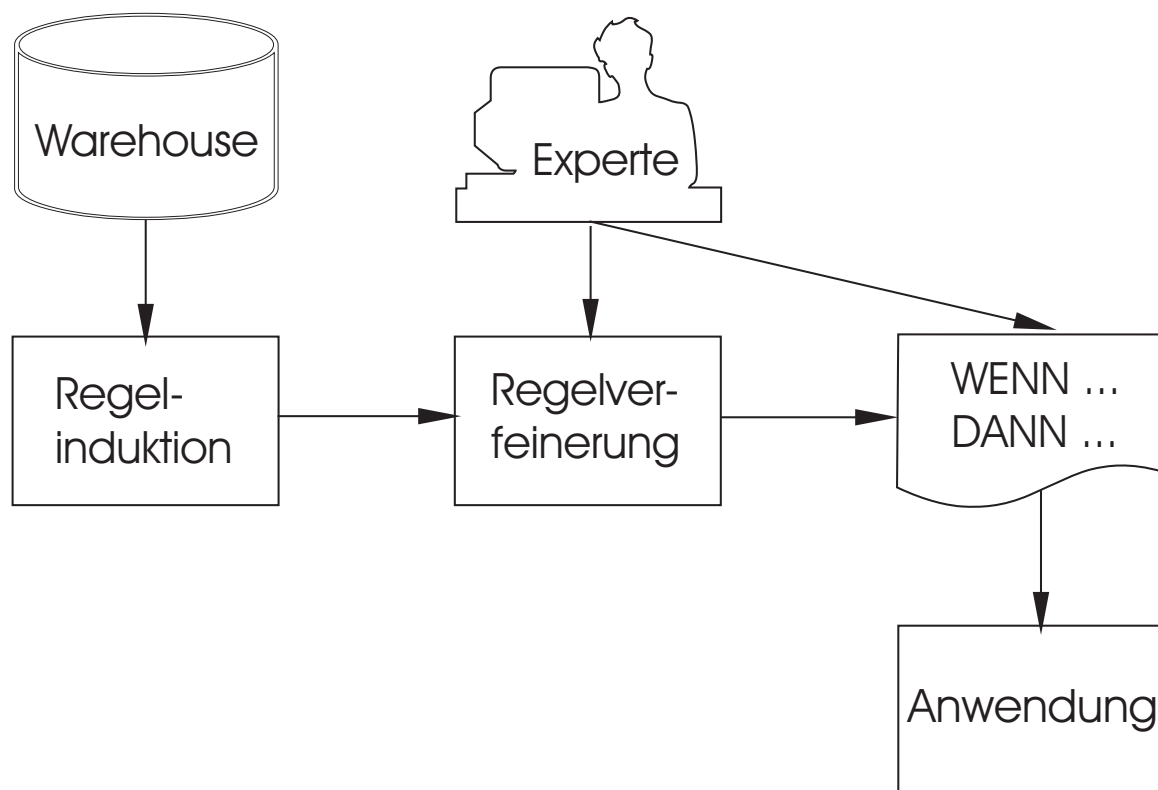
b) *Umgangssprachliche Formulierung?*

Entscheidungsbaum := Menge von hierarchischen Regeln (von Paaren aus **Bedingung** und **Folgerung**)

Die Regelinduktion induziert aus Data Warehouse-Daten Entscheidungsbäume zu gegebenen Forschungsfragen

 Regelinduktion für Entscheidungsbäume

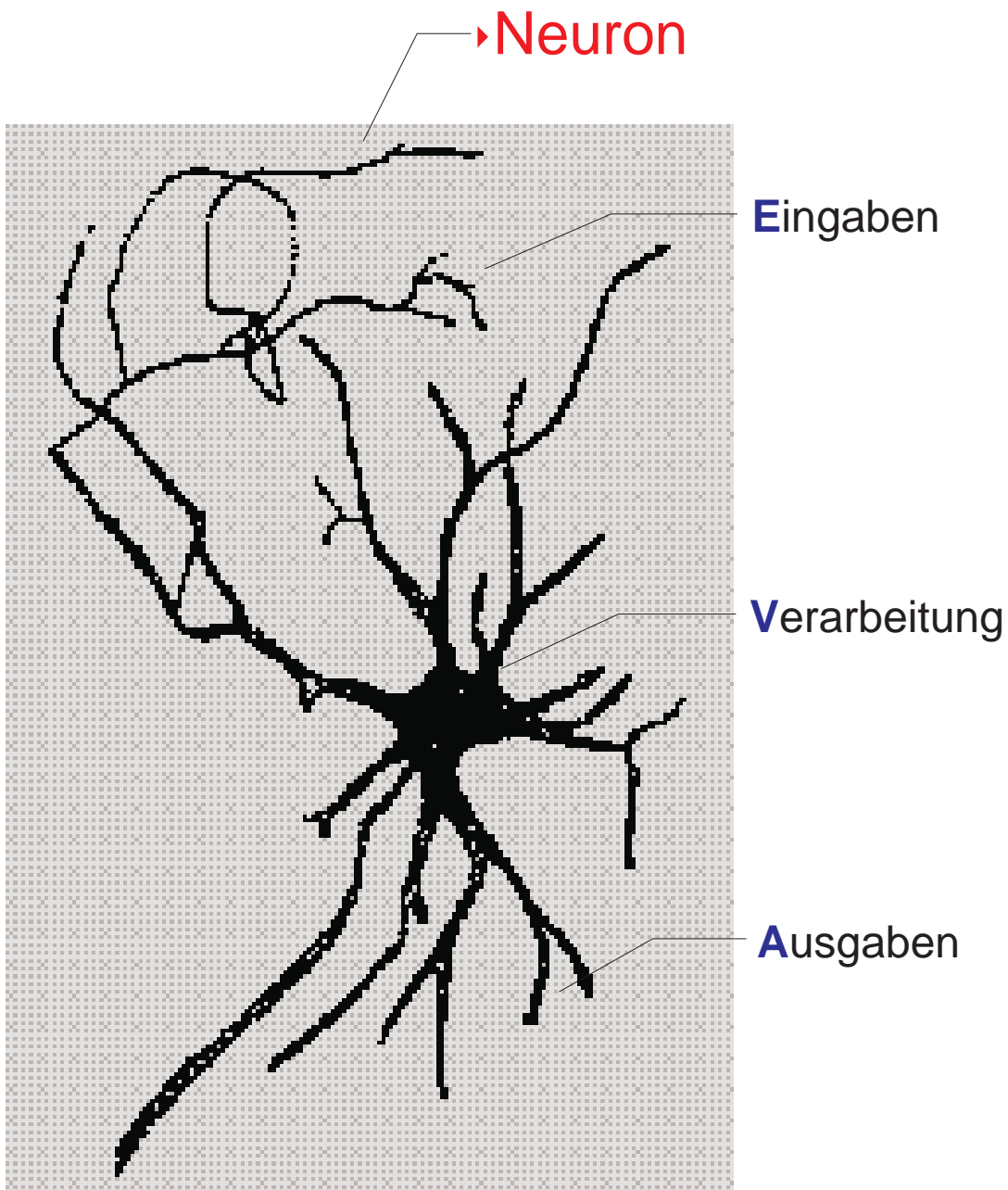
6.6 Induktion von Entscheidungsbäumen



Regelinduktion := Synthese von WENN ... - DANN ... - Regeln aus formatierten Warehouse-Daten →

- ✓ ① Konventionelle Abfragewerkzeuge
- ✓ ② Konventionelle statistische Analyse
- ✓ ③ Induktion von Entscheidungsbäumen (Regeln)
- ⇒ ④ Neuronale Netze
- ⑤ Visualisierung

④ Neuronale Netze

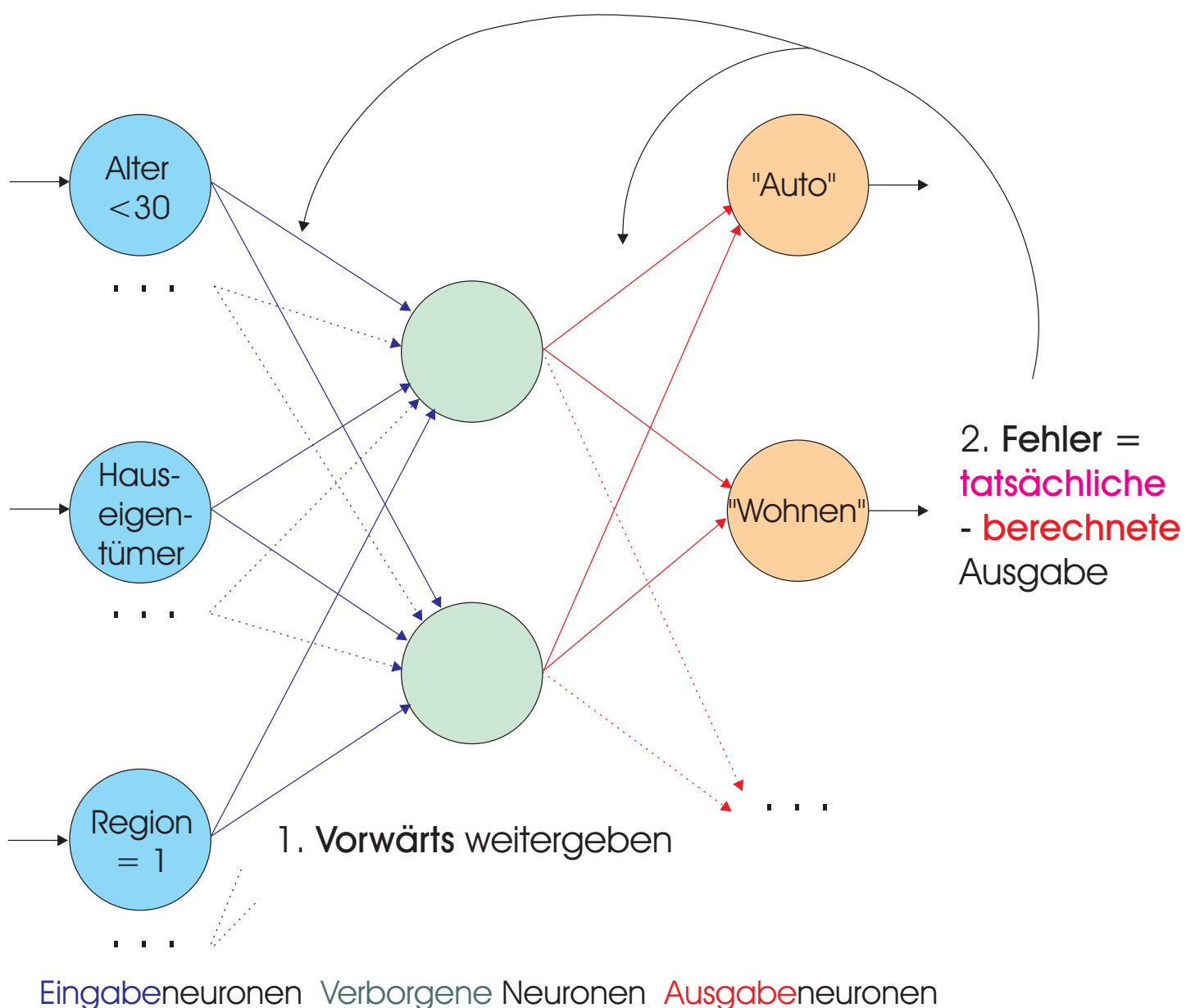


6.7 ZEITSCHRIFTEN - Ein neuronales Netz

Forschungsproblem

Werte von **Eingabeneuronen** (z.B. Alter) solange verarbeiten, bis die Werte der **Ausgabeneuronen** einem gewünschten Kriterium (z.B. den Subskriptionshäufigkeit der Lernmenge) entsprechen

Methode



Einsatz von neuronalen Netzen in ...

Funktionsbereichen wie *Marketing*

- ✓ Kundenklassen identifizieren, die am besten auf bestimmte **Werbeaktionen** ansprechen
- ✓ Neue **Märkte** identifizieren
- ✓ . . .

Branchen wie *Finanzdienstleistungen*

- ✓ Die **Bonität** (Kreditwürdigkeit) eines Antragstellers für einen Konsumkredit vorhersagen →
- ✓ Effiziente **Portfolios** zusammenstellen
- ✓ . . .

Entwicklung neuronaler Netze

Grundgesamtheit definieren



Stichprobe ziehen



Abhängige und Unabhängige vorschlagen



Lernmodell und Software wählen



An der Stichprobe **lernen** und an
einer zweiten Stichprobe **testen**



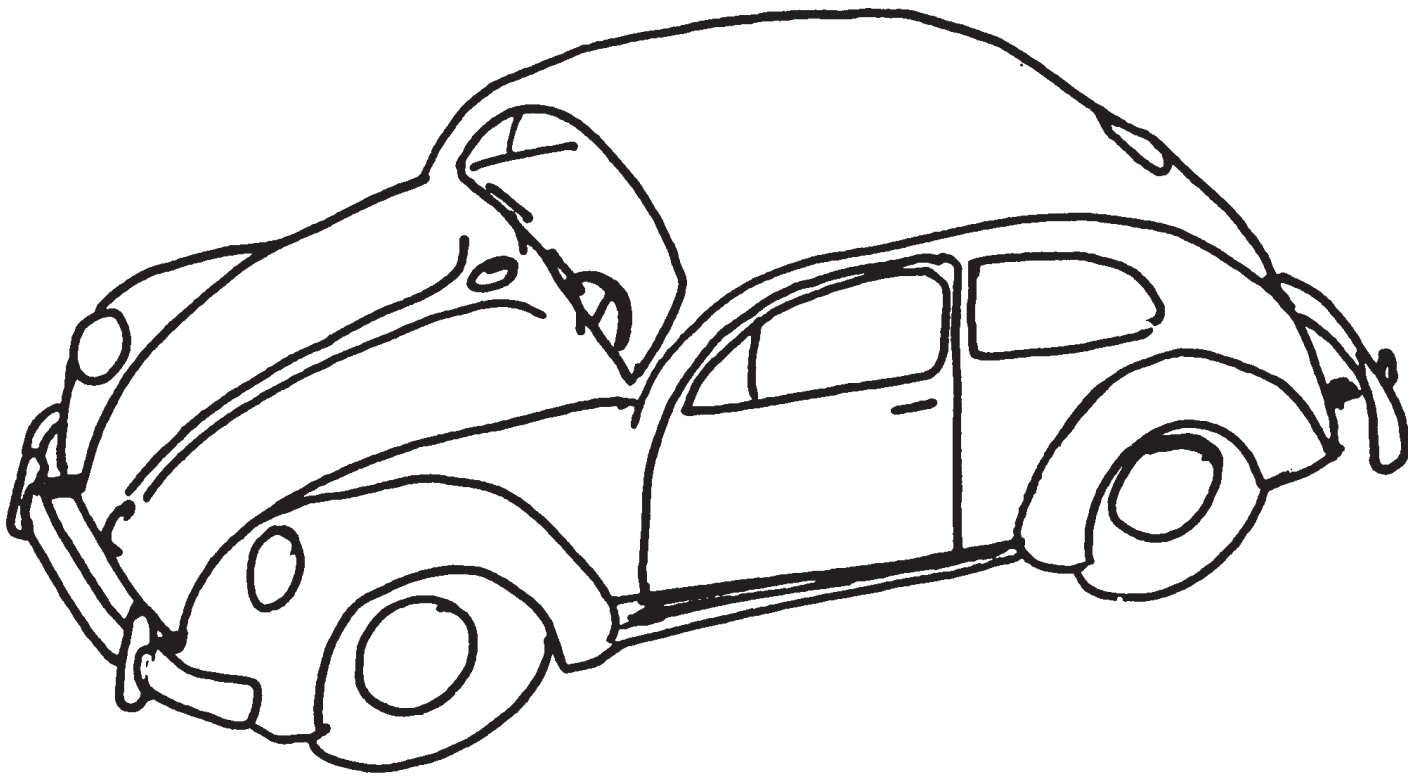
Modell auf neue Daten **anwenden**



Auf die Grundgesamtheit **schliessen**

Unterschiede zur *konventionellen* Statistik ?

⑤ Visualisierung



VW Käfer / FORTRAN:

Er läuft und läuft und ...

A picture is worth a thousand words

- ✓ ① Konventionelle Abfragewerkzeuge
- ✓ ② Konventionelle statistische Analyse
- ✓ ③ Induktion von Entscheidungsbäumen
- ✓ ④ Neuronale Netze
- ⇒ ⑤ Visualisierung

Begriff

Visualisierung :=

- ✓ bildliche Darstellung
- ✓ komplexer Daten zur
- ✓ Entdeckung von Hypothesen oder
- ✓ Veranschaulichung von Ergebnissen

Die Visualisierung ergänzt i.a.
andere Data Mining - Techniken

Einteilung

Diagrammtyp →

...

Dimensionalität →

- Veranschaulichung mehrdimensionaler Abfragen
 - z.B. ›Drilling Down zeigt Details graphisch
 - z.B. ›Drilling Up fasst graphisch zusammen
 - z.B. Animation veranschaulicht ›Slicing and Dicing
- ...

Unterstützte Methode

Veranschaulichung von ...

- ›*OLAP-Ergebnissen*
- Entscheidungsbäumen (›Induktionsverfahren)
- *statistischer Kennzahlen*
 - z.B. von Lage- und Verteilungsparametern
- ...

Diagrammtyp

Kurvendiagramm

Flächendiagramm

Balkendiagramm

engl. bar chart

Kreisdiagramm

engl. pie chart

Streudiagramm →

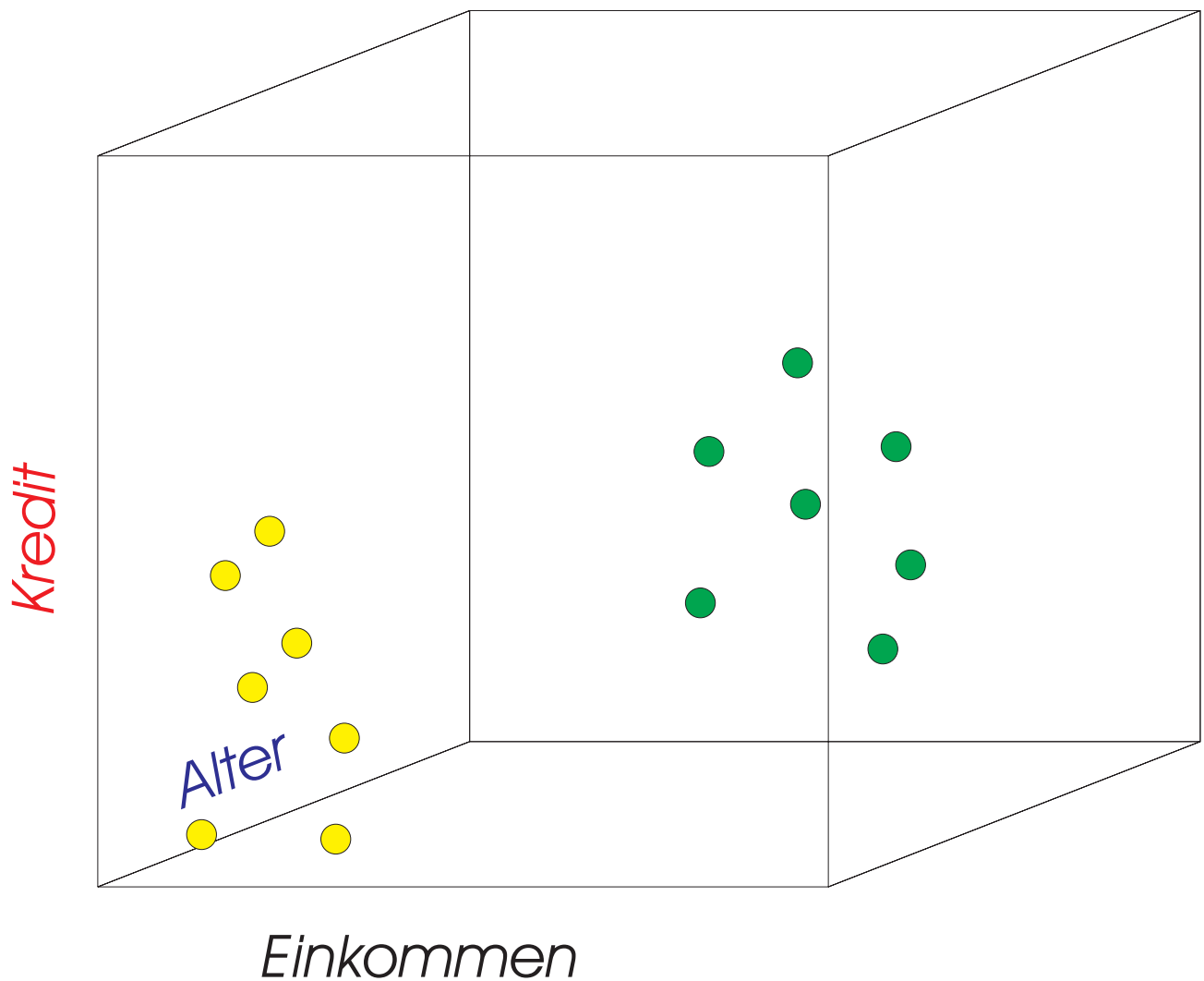
engl. scatterplot

Kartendiagramm

veranschaulicht geographische Daten

. . .

6.8 ZEITSCHRIFTEN - 3D-Streudiagramm



Forschungsproblem

Komplexe Beziehungen veranschaulichen (visualisieren)

Methode

"Musik"	Einkommen	Alter	Kredit
Leser	gering	gering	hoch
Nichtleser	gross	gross	gering

Punkte lassen sich durch drei Koordinaten und zusätzliche Attribute wie Punktfarbe, -grösse, -form, etc. visualisieren

Planung einer ›Direct Mail-Kampagne
für Hunderttausende von Kunden



Auswertung von Kunden mit einer bestimmten
Mindestanzahl letztjähriger Ferngespräche



Visualisierung in einem dreidimensionalen
Scatterplot geometrischer Objekte

X-Achse: Wochentag
Y-Achse: Tageszeit
Z-Achse: Gesprächsdauer
Objektgrösse: Taxkreis
Objektfarbe: ...
Objektform: ...

Weitere Möglichkeiten . . .

- Animation
- Interaktion (Klicks, Drag and Drop, Pivoting, ...)
- Audioeffekte
- Virtuelle Realität
- ...

VISUALISIERUNGSWERKZEUG - Datenauswahl



Untersuchungsziel

Inhaber einer VISA-Karte untersuchen

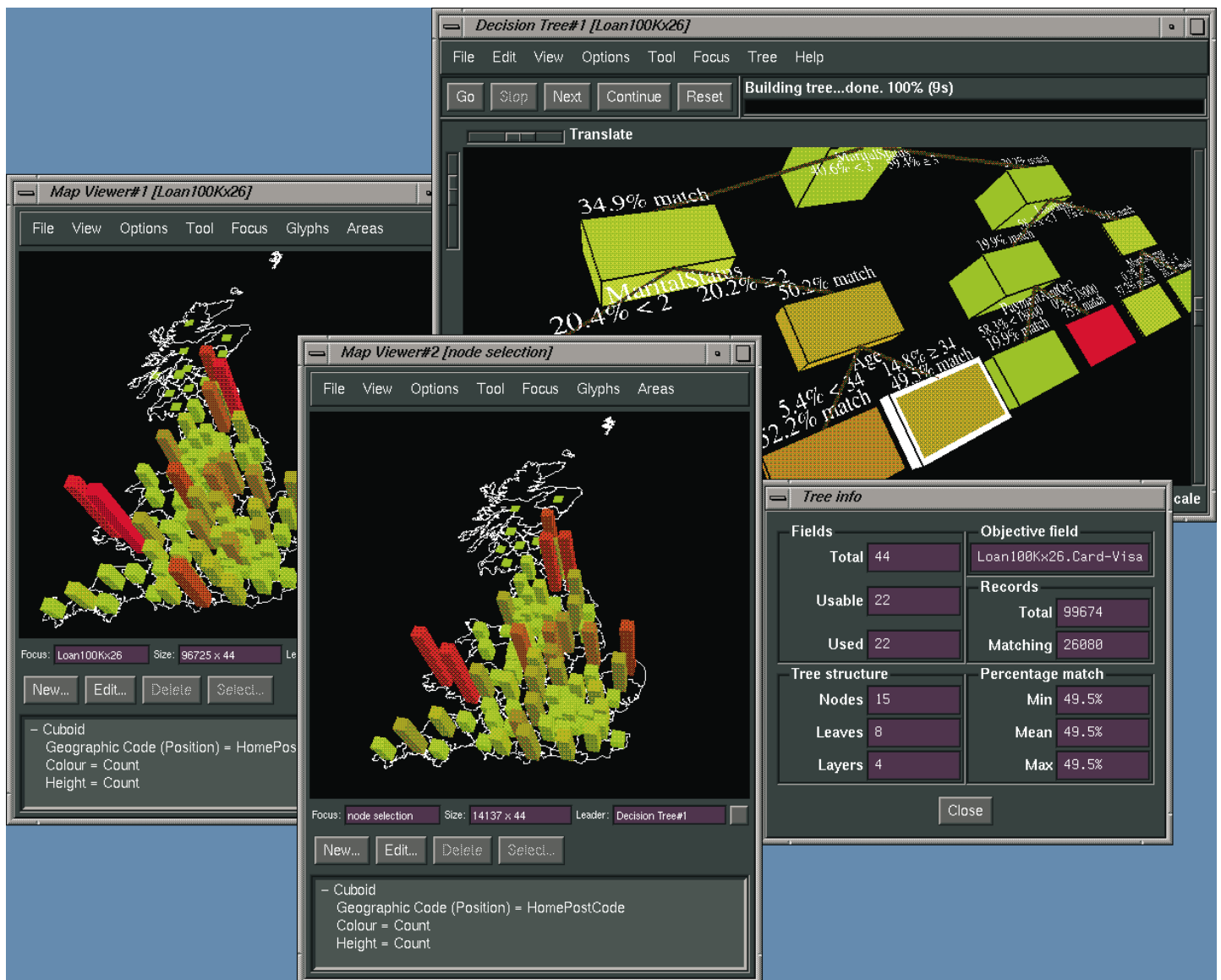
VISUALISIERUNGSWERKZEUG - Methode



Unterstützte Data Mining-Methode ▶ Entscheidungsbaum

- Gesamtstichprobe (Wurzel): 26% VISA
- Rot: 75% vs. 26%, aber kleine Teilstichprobe (0%)
- Orange: zuverlässiger, weil Stichprobe grösser und >26%

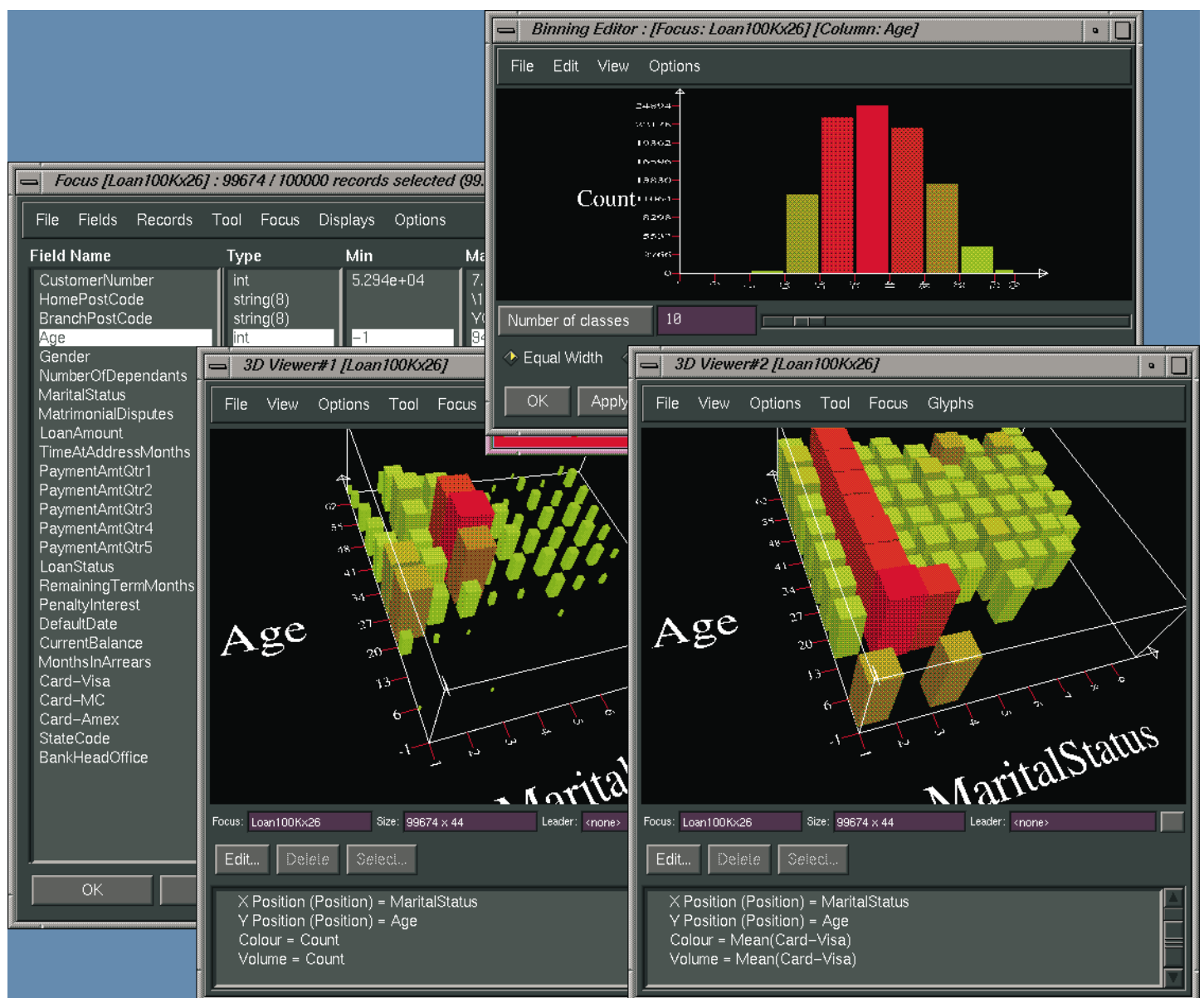
VISUALISIERUNGSWERKZEUG - Dimensionalität



Unterstützte Data Mining-Methode › Drilling

- Drilling Down aus dem › Entscheidungsbaum (weiss)
- Diagrammtyp *Kartendiagramm*

VISUALISIERUNGSWERKZEUG - Diagrammtyp



Diagrammtyp *Balkendiagramm*


- 2D-Balkendiagramm "Häufigkeitsverteilung"
- 3D-Balkendiagramme "Alter vs. Zivilstand"

SPSS Diamond soll Sie in einige Grundfunktionen von Visualisierungswerkzeugen einführen. Im ersten Teil lernen Sie Diamond geleitet kennen. In einem zweiten Teil können Sie das Werkzeug selbständig erkunden.

Lernziele

- ⇒ Grundbegriffe Erhebung, Datensatz, Variable und Wert definieren
- ⇒ Ein-, zwei- und dreidimensionale Beziehungen visualisieren
- ⇒ Farben zur Visualisierung einsetzen

1 SPSS Diamond kennen lernen

- Starten Sie  SPSS Diamond (Warten Sie nach dem Erscheinen des Start-Bildschirms etwa 20 Sekunden auf das Erscheinen der OK-Schaltfläche)
- Wenn Sie den Cursor auf ein Symbol bewegen, erhalten Sie darüber und auf der Statuszeile eine Kurzbeschreibung. Ausführliche Hilfe zum Hauptfenster finden Sie unter dem Menüpunkt *Help/The Main Window* und *Help/Commands*.
- Die Demoversion lässt Sie nur 30 Minuten arbeiten. Starten Sie so oft neu, bis Sie die Aufgabe fertig bearbeitet haben.

a) Hauptfenster (engl. main window)

- Welches sind die Aufgaben der ersten vier Symbole von links?
- Weshalb sind die übrigen Symbole noch deaktiviert?
- Wozu dient das Spreadsheet-Symbol?
- Welche Datenformate kann Diamond importieren?
- Wozu dient die Tabelle mit dem Zeilentitel "Opened Data Set"?

b) Datengesamtheit (engl. data set)

- Laden Sie ...\\diam_le\data\ECONOMY.DA. Was bedeutet die Dateierweiterung .da?

- Wieviele Datensätze (Beobachtungen) und Variablen pro Satz enthält ECONOMY.DA?
- Wenn Sie im Fenster “Selected Data Set Variables” Copy All drücken, so werden die 20 Variablen der Datengesamtheit aktiviert. Da die Demoversion aber höchstens 12 Variablen akzeptiert, müssen Sie 8 Variablen entfernen.
- Interpretieren Sie die nun ausgefüllten ersten vier Zeilen der Tabelle des Hauptfensters.

c) *Rohdaten-Sicht* (engl. raw data view)

- Klicken Sie auf das Symbol “Raw Data” und öffnen Sie ein Fenster mit den Rohdaten von ECONOMY.DA.
- Wo stehen in der Tabelle die Variablen, wo die Beobachtungen?
- Sortieren Sie die Beobachtungen absteigend nach der Variablen Unemployment. Verifizieren Sie das Ergebnis an den Spalten Unemployment und Case #.
- Schliessen Sie das Rohdaten-Fenster.

d) *Bivariate Beziehungen*

- Klicken Sie auf das Symbol “Pairwise” und wählen Sie links die Abszisse Year und rechts die Ordinate DJI (engl. Dow Jones Industrial Average).
- Interpretieren Sie für Year und DJI je die beiden Balkendiagramme. Welche Bedeutung hat ein Balken des einfachen bzw. kumulierten Histogramms.
- Interpretieren Sie das zweidimensionale Streudiagramm.
- Legen Sie eine lineare Regressionsgerade (engl. best-fit straight line) über das Streudiagramm.
- Wie gross ist der Korrelationskoeffizient? Vergleichen Sie den Eindruck, den der numerische Koeffizient vermittelt, mit dem Eindruck des grafischen Streudiagramms.
- Welche Information fehlt zur Beurteilung der Validität des Korrelationskoeffizienten?

e) *Farbliche Visualisierung des Streudiagramms*

- Klicken Sie auf das Symbol “Create Red Region”.
- Bewegen Sie die Maus auf den Punkt mit der Abszisse 1982. Drücken Sie dann die linke Maustaste und zeichnen Sie ein Rechteck bis zur rechten oberen Ecke des Streudiagramms. Sie markieren so einen Teil des Streudiagramms, der eine Regressionsgerade mit einem höheren Korrelationskoeffizient ergibt.
- Vergleichen Sie die rote Regressionsgerade mit der ursprünglichen (Falls keine rote Gerade erscheint, klicken Sie auf das mittlere der drei Schaltflächen der Ergebnistabelle). Wie gross ist der Korrelationskoeffizient der roten Gerade?
- Experimentieren Sie mit verschiedenen Positionen des Rechtecks. Verschieben Sie das Rechteck, indem Sie Alt drücken und die Maus bei niedergedrückter Linkstaste bewegen. Die Abmessungen des Rechtecks ändern Sie mit Ctrl statt Alt.
- Markieren Sie den Rest des Streudiagramms mit einem blauen Rechteck und interpretieren Sie Regressionsgerade und den Korrelationskoeffizienten.

f) *Gemeinsame Visualisierung mehrerer Streudiagramme*

- Klicken Sie auf das Symbol “Directory”. Suchen Sie das Streudiagramm “DJI - Year”.
- Mit der Kompassrose ändern Sie die Position der Directory-Elemente. Durch einen Doppelklick auf ein beliebiges Streudiagramm können Sie das Diagramm näher untersuchen. Klicken Sie zum Beispiel auf “BondLong - Trade”. Was bedeuten die Variablen BondLong und Trade?

g) *Zwei- und dreidimensionale Animation eines Streudiagramms*

Ein parametrisches “Schlangendiagramm” ist ein zweidimensionales Streudiagramm, dessen Punkte in Abhängigkeit von einer dritten (parametrischen) Variable verbunden werden.

- Klicken Sie auf das Symbol “Parametric Snake”.

- Wählen Sie die Abszisse Trade, die Ordinate BondLong und die parametrische Variable Year.
- Wenn Sie auf Start klicken, dann werden die Punkte in aufsteigender Reihenfolge der Werte der parametrischen Variablen (hier Year) verbunden.

Eine weitere Art der Animation ist das dreidimensionale Streudiagramm.

- Klicken Sie auf das Symbol “Triplewise”.
- Experimentieren Sie mit den Schaltflächen “Start”, “Adjust Orientation” und “Adjust View”.

h) *Visualisierung von Korrelationen*

Ein Fractal Foam Window veranschaulicht uni- oder bivariate Statistiken. In der Mitte liegt die gewählte Fokusvariable als weisse Blase (engl. bubble). Die best korrelierten Variablen liegen im Gegenurzeigersinn um die Fokusvariable herum. Je grösser der absolute Korrelationskoeffizient, desto grösser die Blase. Jede Blase kann ihrerseits (rekursiv) Blasen um sich ordnen.

- Klicken Sie auf das Symbol “Fractal Foam”.
- Wählen Sie die Variable BondLong.
- Bewegen Sie die Maus über die weisse Bubble.
- Welches ist die Variable, die mit BondLong am stärksten korreliert ist? Klicken Sie doppelt auf diese Variable.
- Weshalb ist es gefährlich, bivariate Korrelationen als Masse für den Zusammenhang zwischen zwei von mehreren Variablen zu interpretieren, wenn die Variablen alle unter sich verbunden sind?

2 *Diamond selbständig anwenden*

Versuchen, Sie über die Hilfedatei und freies Experimentieren die folgenden Konzepte zu begreifen:

- Quadwise Window

- Univariate und bivariate Statistiken
- Transformation und Definition von Variablen
- Teilmengenbildung (engl. reinvocation)
- Import von Fremdformaten (zum Beispiel einer MS Excel-Tabelle)

Data Mining-Werkzeug

Data Mining-Werkzeug := Software, die ...

- ✓ Daten so **vorbereitet**, dass sie ...
- ✓ Ergebnisse mit Data Mining-Methoden **berechnen** ...
- ✓ und **präsentieren** kann



Werkzeuge beschränken sich auf die
Ausführung von Data Mining-Algorithmen

6.9 Methoden und Werkzeuge im Überblick

<i>Werkzeugklassen</i>	<i>Methodenbeispiele</i>	<i>Werkzeugbeispiele</i>
Abfragesprachen und Berichtsgeneratoren	›QBE- und ›SQL-Frontends	MS Access Cognos Impromptu
Tabellenkalkulation	← ›Was-Wenn-Analyse	MS Excel
Multidimensionale Abfragewerkzeuge	← ›OLAP	DecisionSuite von Information Advantage Cognos PowerPlay Synchrony
① Konv. Statistik	Entdecken , z.B. ›Clusteranalyse ›Faktoranalyse Testen und Schliessen , z.B. ›Varianzanalyse ›Regressionsanalyse ›Diskriminanzanalyse	Kern von SPSS statist. Teil von SAS
② Induktion von Entscheidungsbäumen	›ID3-Algorithmus	XpertRule Profiler Cognos Scenario
③ Neuronale Netze und Genetische Algorithmen	›Neuronale Netze →	Predict
④ Visualisierung	Visuelle ›Sensitivitätsanalyse	AVS / Express IBM Visual Explorer

Konventionelle Analyse

Data Mining im **engeren** Sinn

Data Mining i. **weiteren** Sinn

Data Mining i. **weitesten** Sinn

6.10 Werkzeugkriterien

Anwendungsphase	Beurteilungskriterium
Eingabe	Datenbank Einzeltabelle mehrere Tabellen
	Tabellenkalkulationsblätter
	Textdateien
	Gedrucktes z.B. ›OCR-Schnittstelle
Verarbeitung	Datenaufbereitung Mischen Trennen Gruppen bilden (Aggregation) Integrität prüfen Transformieren Fehlende Werte verwalten
	Methoden Zahl Umfang der Unterstützung
	Entwicklungsphasen
	Komponenten-Schnittstelle z.B. ›COM z.B. ›ActiveX
Ausgabe	Ad hoc-Analyse durch den Endbenutzer
	Makrosprache
	Grafik World Wide Web

Wenige Methoden und viele Werkzeuge

Marketing-Argumente ...

Unsere Werkzeuge und Methoden sind ...

- ☺ brandneu
- ☺ originell
- ☺ kombiniert

... und die Wirklichkeit

- ☹ alter Wein in neuen Schläuchen
- ☹ proprietäre Methoden nicht nachvollziehbar
- ☹ Methodenkombinationen schwer verständlich

Statistik-Produkte

SAS (...)

- Datenvorbereitung und -ausgabe
- Data Warehouse-Verbindung
- Konventionelle Statistik
- ›Visualisierung
- Beratung

Stärke: Vorbereitung von Massendaten

SPSS (Statistical Package for the Social Sciences)

- Dateneingabe und -ausgabe
- Konventionelle Statistik
- ›Entscheidungsbaum (›CHAID, ›CART)
- ›Neuronale Netze
- ›Visualisierung

Stärke: Einbezug neuerer Verfahren wie neuronale Netze

MathSoft

- S-Plus

Stärke: Erweiterbarkeit

Ein typisches Paket zur konventionellen Statistik

Deskriptive Statistiken

z.B. Lage- und Verteilungsmasse

Vergleich von Mittelwerten

z.B. ›t-Test

›ANOVA (Analysis of Variance)

Korrelation

z.B. einfache lineare ›Korrelation

›Regression

z.B. lineare ›Mehrfachregression


Nichtparametrische Tests

z.B. Chi-Quadrat-Test

Weitere multivariate Verfahren

- ›Faktorenanalyse
- ›Clusteranalyse
- ›Diskriminanzanalyse

...

Ein stark vereinfachtes Datenmodell und das entsprechende MS Access-Beispiel  [Bank.mdb](#) beschreiben operative und analytische Daten einer Bank. Anhand von zwei Fragestellungen lernen Sie die Data Mining-Phasen Problem- und Modellspezifikation kennen.

Falls Sie direkt von der CD ROM oder vom Server laden, können Sie in der Regel nicht auf die geladenen Dateien schreiben. Kopieren Sie deshalb unmittelbar vor dem Laden die Datei auf Ihre Festplatte (z.B. D:\) und entfernen Sie den Schreibschutz.

① **Kreditkartenbetrug**

Eine zuverlässige Identifikation betrügerischer Kreditkarten-Transaktionen spart Banken und Kreditkarten-Organisationen Kosten. Beantworten Sie dazu die folgenden Fragen:

- a) Welche *Attribute* des folgenden Datenmodellausschnitts eignen sich zur Klassifikation oder Vorhersage betrügerischer Transaktionen?
- b) Erweitern Sie den Datenmodell-Ausschnitt mit *zusätzlichen* Attributen, die sich zur Identifikation von Kreditkartenbetrug eignen.
- c) Viele Data Mining-Werkzeuge lassen sich nicht so gut in Datenbanksysteme integrieren, dass die bloße Auswahl relevanter Attribute und Beziehungen aus einem Datenbankschema genügt.

Integrieren Sie deshalb die unter a) ausgewählten Attribute in eine *temporäre Tabelle* oder eine Sicht (*View*), welche das Data Mining-Werkzeug direkt verwenden kann. Nennen Sie die notwendige(n) SQL-Anweisung(en) (Verwenden Sie als laufendes Datum den 1/7/97).

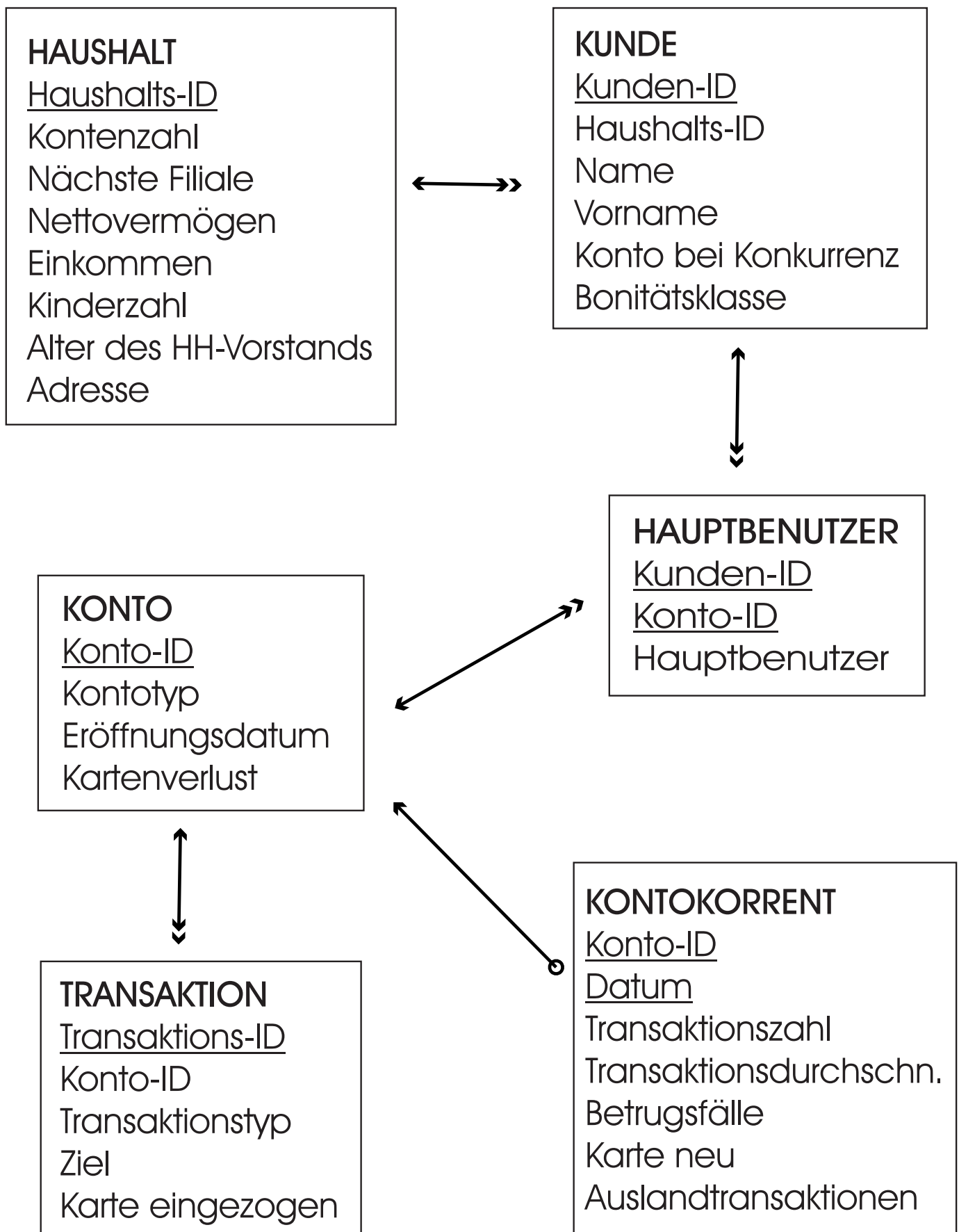
- d) Welche der bisher erwähnten Data Mining-*Methoden* könnten sich für die Analyse eignen?

② Retention Management

In umkämpften Märkten kann es gewinnbringender sein, bestehende Kunden zu erhalten, als neue zu gewinnen. Eine Bank kann deshalb versuchen, Kundenwechsel frühzeitig zu erkennen, um entsprechende Massnahmen einzuleiten (Retention Marketing).

- a) Welche *Attribute* eignen sich zur Vorhersage von Kundenwechseln?
- b) Erweitern Sie den Datenmodell-Ausschnitt mit *zusätzlichen* Attributen, die sich zur Vorhersage der Kundenfluktuation eignen.
- c) Welche der bisher erwähnten Data Mining-*Methoden* eignen sich für die Analyse.
- d) Eine abnehmende Zahl von Transaktionen eines Kunden kann ein Indiz für einen bevorstehenden Bankenwechsel sein. Sammeln Sie deswegen für alle Kunden, die bereits ein Konto bei der Konkurrenz haben, die Zahl der Transaktionen der letzten sechs Monate. Nennen Sie die dazu notwendigen *SQL-Anweisungen* (Verwenden Sie als laufendes Datum den 1/7/97).

Ausschnitt aus dem Datenmodell (A 6.1)



Data Mining auf Tabellenblättern?

☹ Zahl und Funktionalität der Methoden

- ›Entscheidungsbäume
- ›Neuronale Netze
- ›Zeitreihenanalyse
- Qualitätskontrolle
- nichtparametrische Tests

☹ Grosse Datenmengen

☹ Eingabeprüfungen

☹ Ausgabe

Funktionsumfang der Grafikausgabe

- ›Mehrdimensionalität

☹ Verwaltung von Metainformation

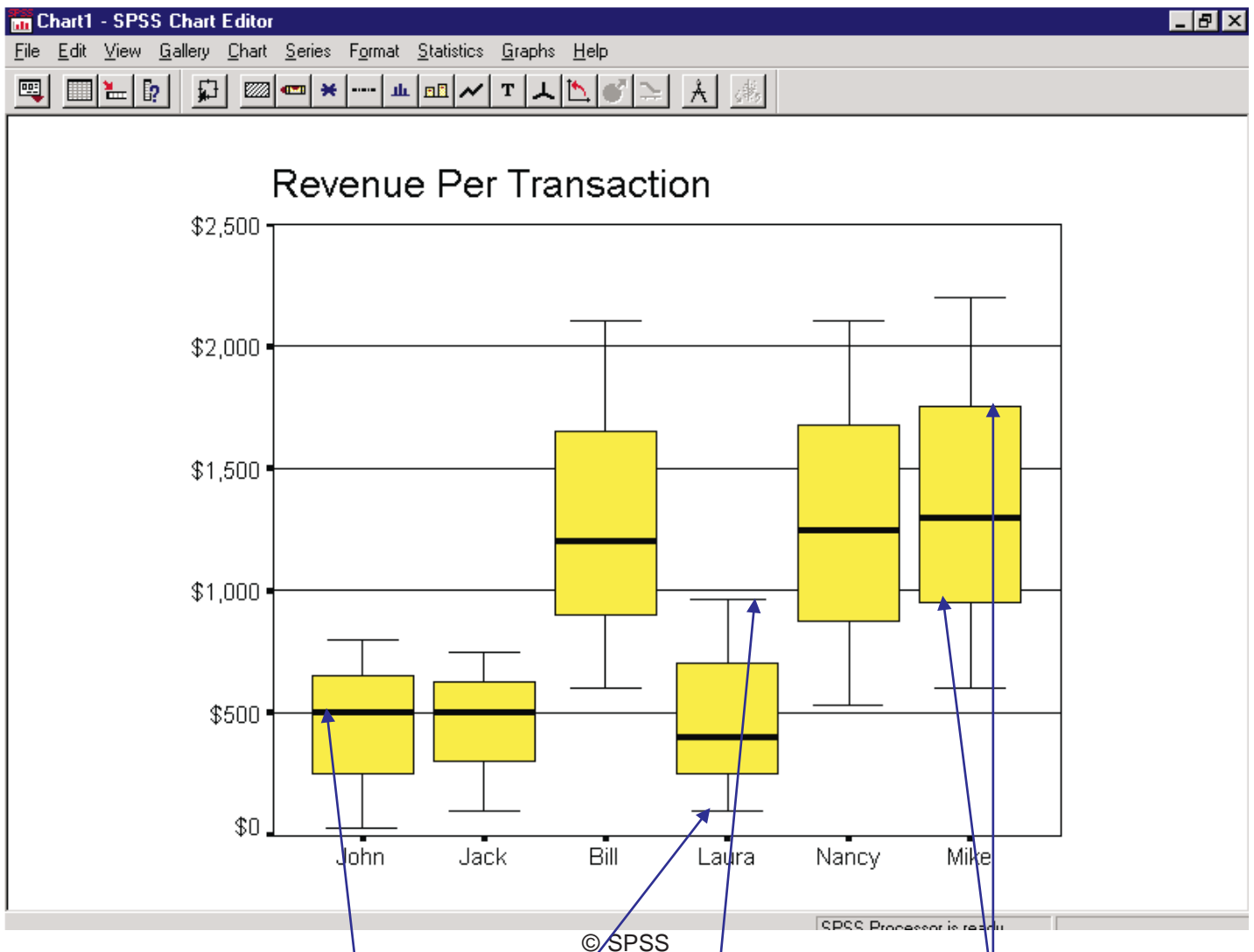
z.B. ›Data Dictionary

☹ Verwaltung fehlender Werte

Typisierung fehlender Daten

Ausschluss fehlender Daten von Berechnungen

Boxplot



Rechteck (engl. box) zwischen unterem und oberem Quartil

Querstrich beim Median

Querstriche beim Minimum und Maximum

Spreadsheet-Lücke Multidimensionalität

Untitled - SPSS Output Navigator

File Edit View Insert Format Statistics Graphs Utilities Window Help

Type of vacation Self-planned trip

		Gender	
		Male	Female
		Col %	Col %
How enjoyable was vacation?	Not Enjoyable at all		.9%
	Not Too Enjoyable	5.1%	1.7%
	O.K.	11.0%	5.2%
	Enjoyable	48.5%	50.0%
	Very Enjoyable	35.3%	42.2%

SPSS Processor is ready

Untitled - SPSS Output Navigator

File Edit View Insert Format Statistics Graphs Utilities Window Help

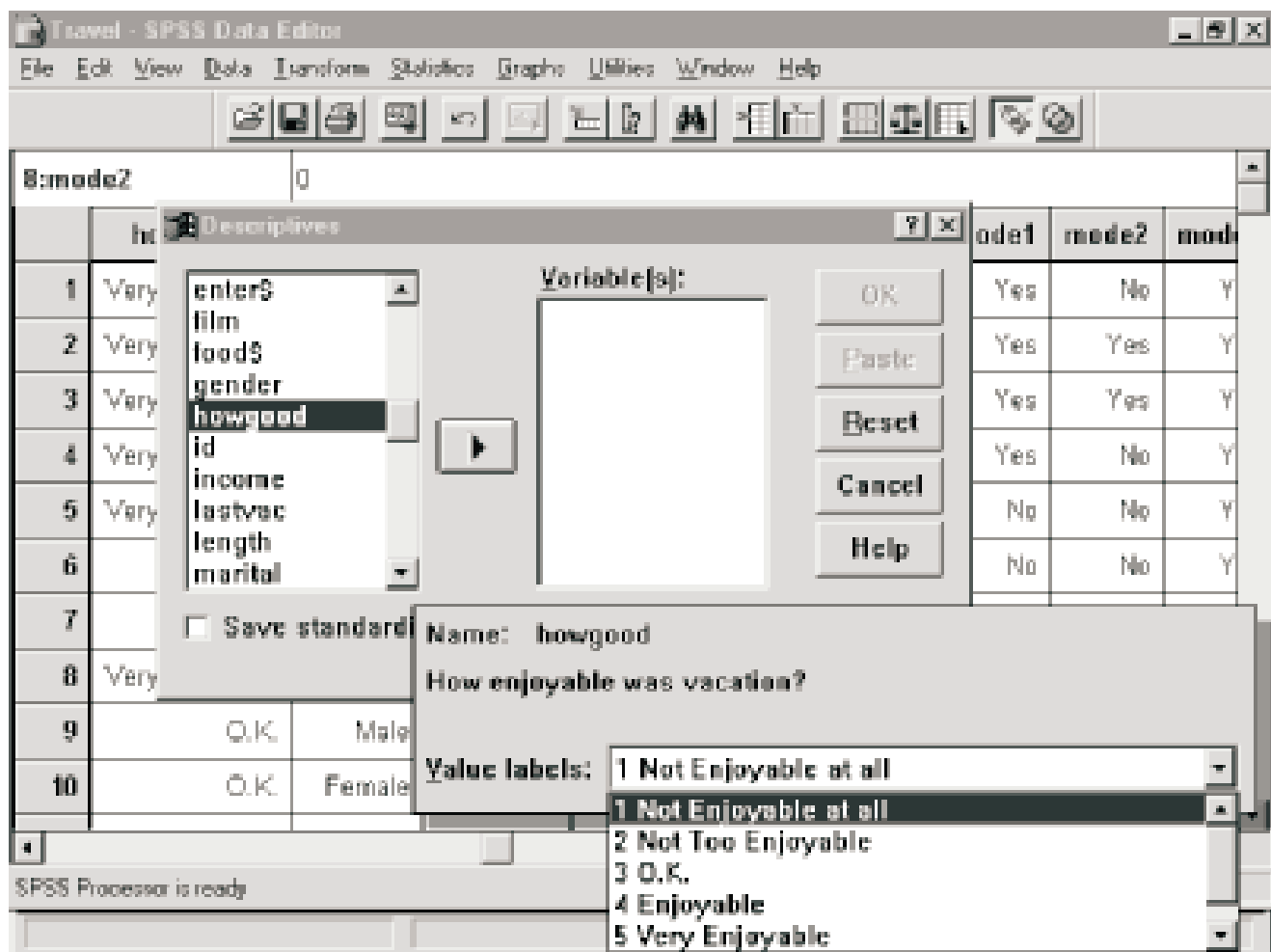
Type of vacation Packaged holiday

		Gender	
		Male	Female
		Col %	Col %
How enjoyable was vacation?	Not Enjoyable at all	7.7%	.0%
	Not Too Enjoyable	.0%	.0%
	O.K.	11.5%	.0%
	Enjoyable	61.5%	68.8%
	Very Enjoyable	19.2%	31.3%

SPSS Processor is ready

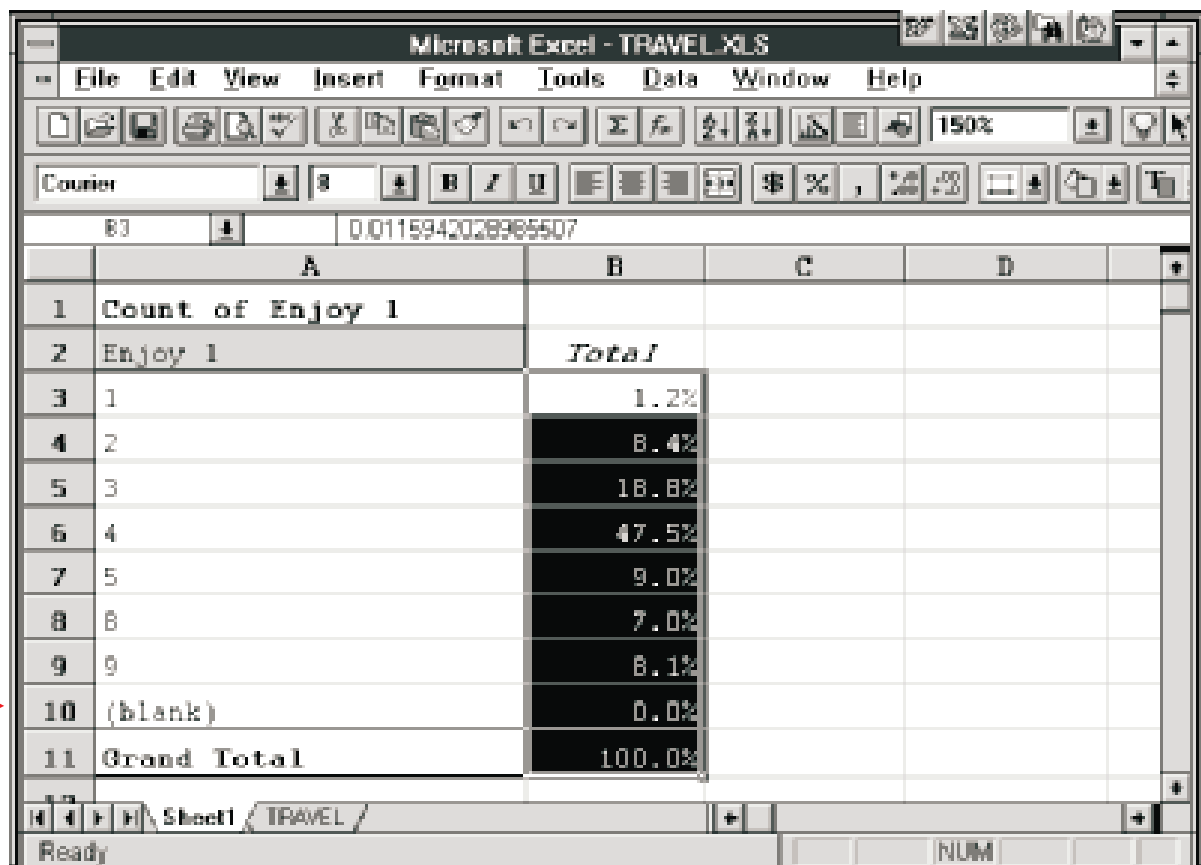
Drei Dimensionen

Spreadsheet-Lücken Metadaten verwalten



© SPSS

Spreadsheet-Lücke Fehlende Werte



Microsoft Excel - TRAVEL.XLS

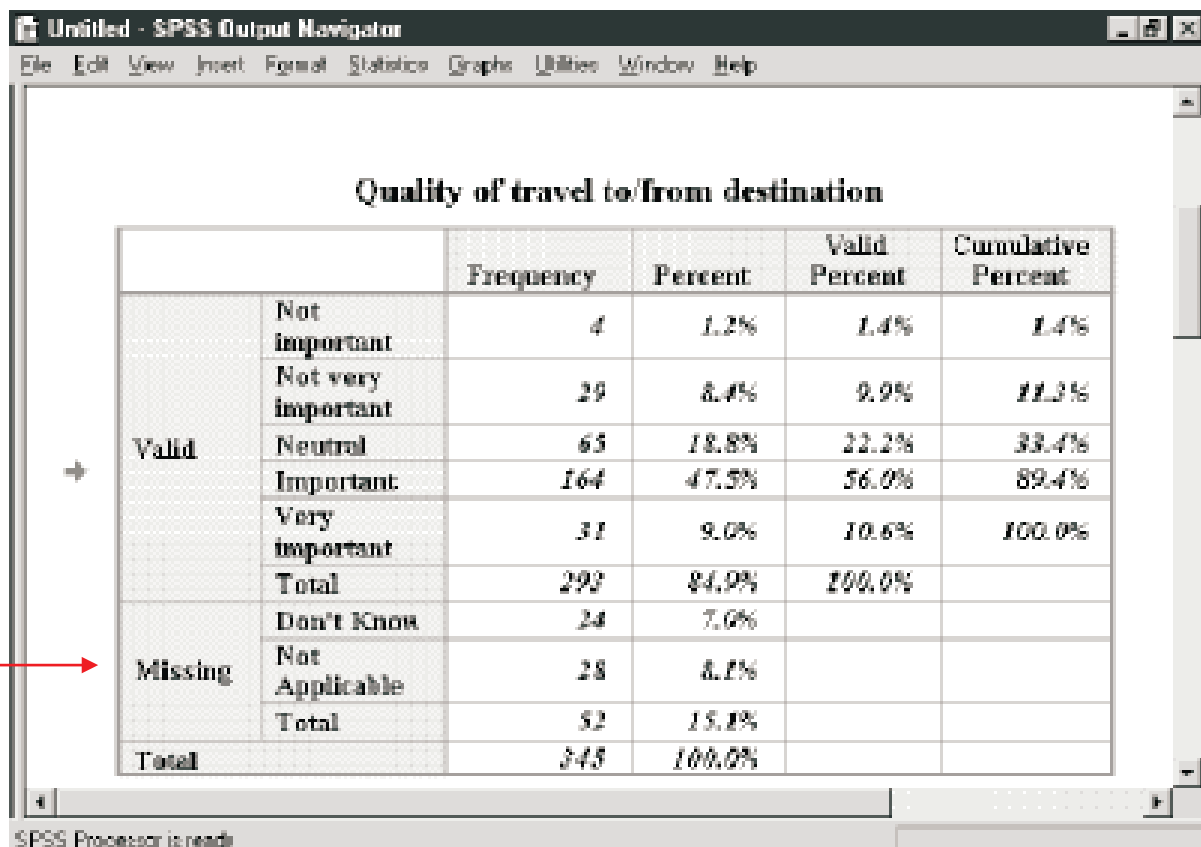
File Edit View Insert Format Tools Data Window Help

Counter 83 0.0115942028985507

	A	B	C	D
1	Count of Enjoy 1			
2	Enjoy 1	Total		
3	1	1.2%		
4	2	8.4%		
5	3	18.8%		
6	4	47.5%		
7	5	9.0%		
8	6	7.0%		
9	9	8.1%		
10	(blank)	0.0%		
11	Grand Total	100.0%		

Ready NUM

© SPSS



Untitled - SPSS Output Navigator

File Edit View Insert Format Statistics Graphs Utilities Window Help

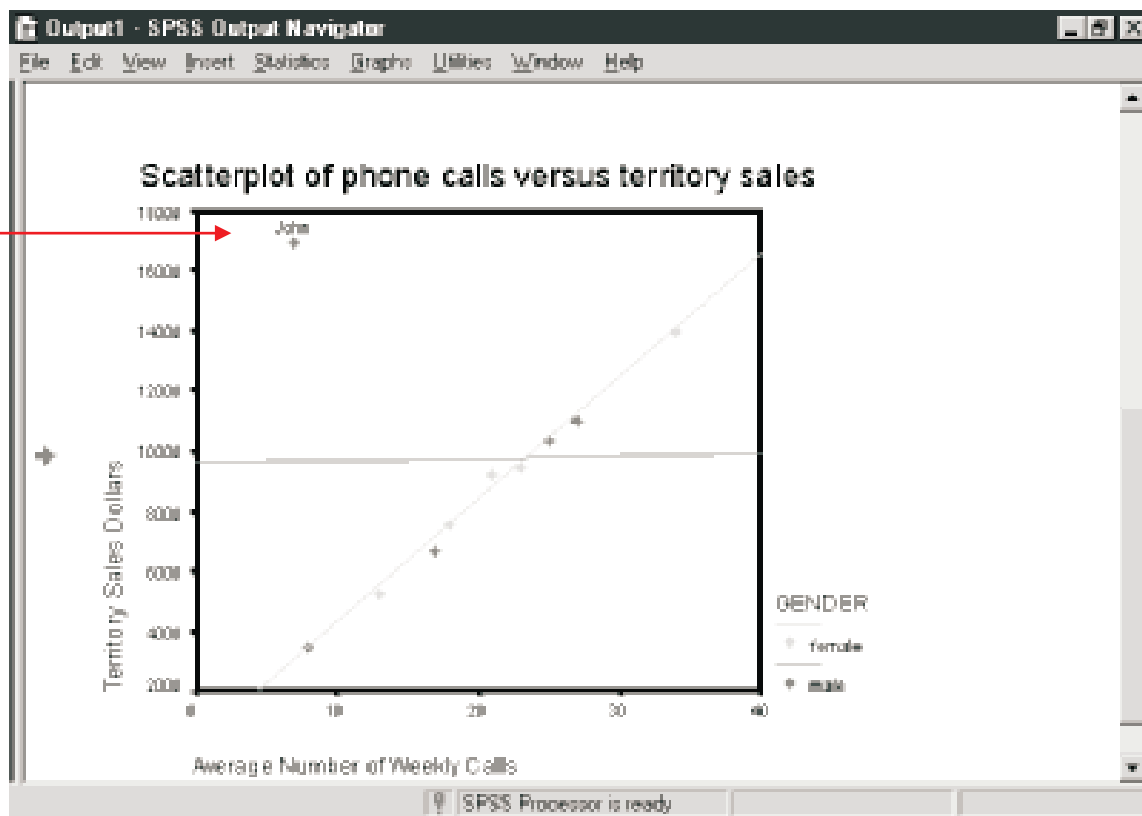
Quality of travel to/from destination

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not important	4	1.2%	1.4%	1.4%
	Not very important	29	8.4%	9.9%	11.3%
	Neutral	65	18.8%	22.2%	33.4%
	Important	164	47.5%	56.0%	89.4%
	Very important	31	9.0%	10.6%	100.0%
	Total	293	84.9%	100.0%	
Missing	Don't Know	24	7.0%		
	Not Applicable	28	8.1%		
	Total	52	15.1%		
Total		345	100.0%		

SPSS Processor is ready

© SPSS

Spreadsheet-Lücke Eingabefehler beheben



© SPSS

The figure shows the SPSS Data Editor window with a table containing 10 rows of data. The columns are salesrep, calls, termdol\$, gender, and three empty columns labeled var. Row 7, corresponding to John, is highlighted in black. A red arrow points to this row. The SPSS Processor is ready.

	salesrep	calls	termdol\$	gender	var	var	var
1	Bob	25	\$10,367	male			
2	Mary	21	\$9,264	female			
3	Jane	34	\$13,976	female			
4	Robert	17	\$6,735	male			
5	Stacey	13	\$5,266	female			
6	Susan	23	\$9,463	female			
7	John	7	\$18,928	male			
8	Larry	8	\$3,512	male			
9	Joan	18	\$7,809	female			
10	Mark	27	\$11,078	male			

© SPSS

Spreadsheet-Lücke Komfortables Pivoting

Customer recommendation by store

Contact with an employee			Would you recommend the store?				
			Extremely likely	Very likely	Somewhat likely	Not very likely	Not at all likely
Yes	Store location	Norfolk	23%	49%	15%	10%	3%
		Virginia Beach	21%	46%	31%	3%	0%
		Chesapeake	24%	39%	29%	5%	3%
		Portsmouth	39%	29%	23%	10%	0%
		Suffolk	18%	61%	11%	7%	2%
	Total		24%	46%	21%	7%	2%
No	Store location	Norfolk	40%	20%	20%	0%	20%
		Virginia Beach	44%	56%	0%	0%	0%
		Chesapeake	20%	33%	27%	13%	7%
		Portsmouth	37%	26%	26%	0%	11%
		Suffolk	0%	100%	0%	0%	0%
	Total		33%	33%	20%	4%	9%

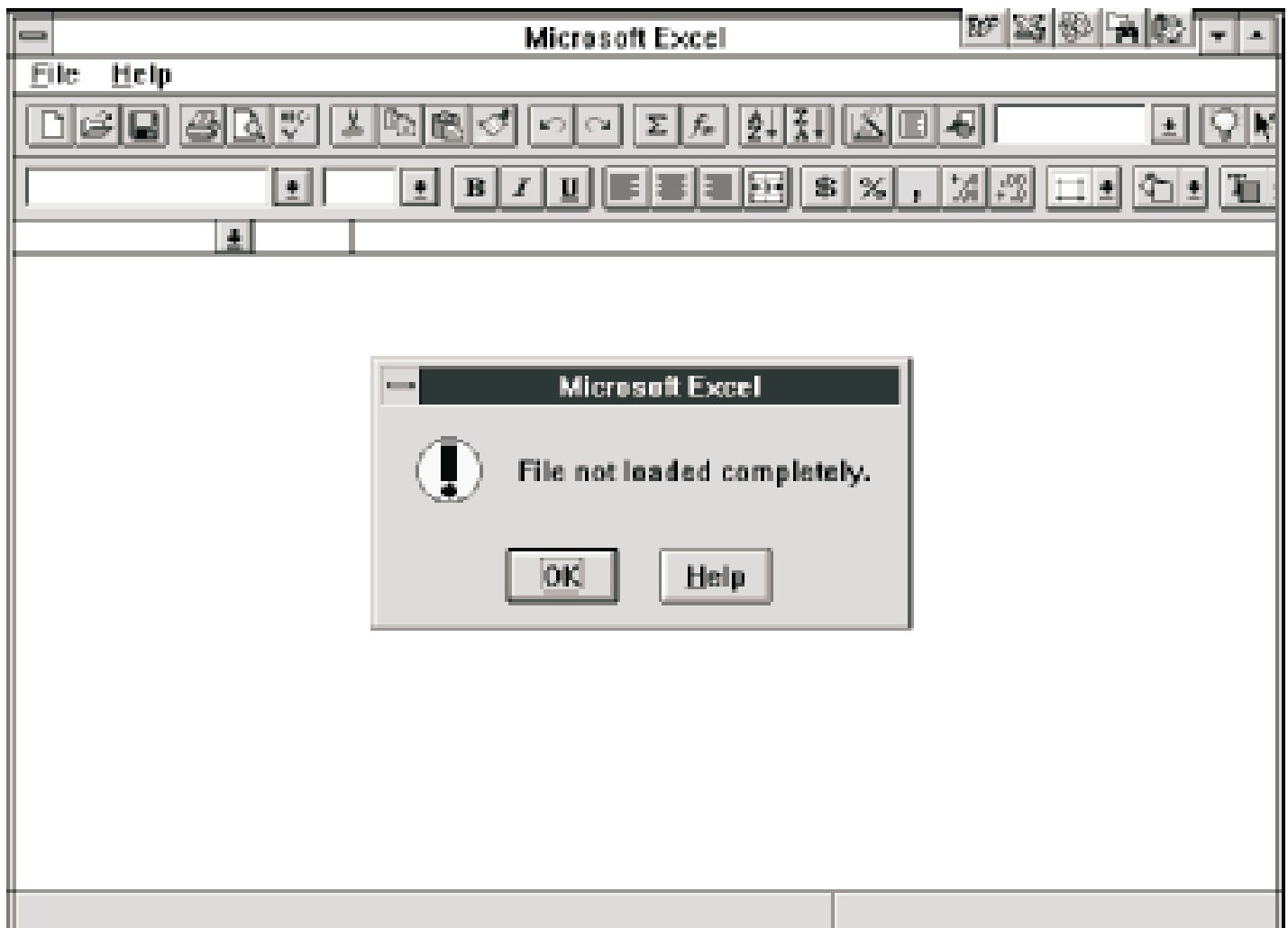
© SPSS

Customer recommendation by store

		Contact with an employee									
		Yes					No				
		Would you recommend the store?					Would you recommend the store?				
		Extremely likely	Very likely	Somewhat likely	Not very likely	Not at all likely	Extremely likely	Very likely	Somewhat likely	Not very likely	Not at all likely
Store location	Norfolk	23%	49%	15%	10%	3%	40%	20%	20%	0%	20%
	Virginia Beach	21%	46%	31%	3%	0%	44%	56%	0%	0%	0%
	Chesapeake	24%	39%	29%	5%	3%	20%	33%	27%	13%	7%
	Portsmouth	39%	29%	23%	10%	0%	37%	26%	26%	0%	11%
	Suffolk	18%	61%	11%	7%	2%	0%	100%	0%	0%	0%
Total		24%	46%	21%	7%	2%	33%	33%	20%	4%	9%

© SPSS

Spreadsheet-Lücke **Grosse Datenmengen**



Tabellenkalkulations- und Datenbanksoftware
genügt nur für einfache Auswertungen !

Internet-Adressen

Produktunabhängige Information

<http://www.kdnuggets.com/> (The Knowledge Discovery Mine)

Umfangreicher Überblick. Abonnement möglich

<http://www.andypryke.com/university/TheDataMine.htm> (The Data Mine)

<http://www.wkap.nl/journalhome.htm/1384-5810> (Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers Group)

<http://www.voyager.net/uthed/neural.htm> (Neuronale Netze)

Produktübersichten

<http://www.dwinfocenter.org/datamine.html>

<http://www.twocrows.com> (Data Mining Software)

<http://www.cutter.com> (Software allgemein)

<http://www.yphise.com> (Software allgemein)

(Verweise auf Anbieter finden Sie auf den betreffenden Folien)

Produktunabhängige News Groups

<http://www.emsl.pnl.gov:2080/docs/cie/neural/newsgroups.html>

Neuronale Netze

 [Web Quiz](#)

Folienverzeichnis (Ein Klick führt zur gewünschten Folie)

<u>6.1 Data Mining</u>	<u>2</u>
<u>Einordnung</u>	<u>3</u>
<u>Einblick in das Data Mining</u>	<u>4</u>
<u>Data Mining</u>	<u>5</u>
<u>6.2 Anwendungen</u>	<u>6</u>
<u>Anwendungsphasen</u>	<u>7</u>
<u>6.3 Anwendungsklassen</u>	<u>8</u>
<u>Methodenklassen</u>	<u>9</u>
<u>Fussangel des Data Mining</u>	<u>10</u>
<u>Voraussetzungen des Data Mining</u>	<u>11</u>
<u>Data Mining ist meist datengetrieben</u>	<u>12</u>
<u>6.4 Daten als Ausgangspunkt</u>	<u>13</u>
<u>Entwicklungsphasen</u>	<u>14</u>
<u>① Data Mining im weiteren Sinne</u>	<u>15</u>
<u>6.5 📌 ZEITSCHRIFTEN - Data Mining i.e.S.</u>	<u>16</u>
<u>📌 Naive Vorhersage</u>	<u>17</u>
<u>📌 Methodische Vorhersage</u>	<u>18</u>
<u>② Klassifikation und Clustering in der Statistik</u>	<u>19</u>
<u>③ Entscheidungsbaum</u>	<u>20</u>
<u>6.6 Induktion von Entscheidungsbäumen</u>	<u>21</u>
<u>④ Neuronale Netze</u>	<u>22</u>
<u>6.7 ZEITSCHRIFTEN - Ein neuronales Netz</u>	<u>23</u>
<u>Anwendungen neuronaler Netze</u>	<u>24</u>

Entwicklung neuronaler Netze	25
⑤ Visualisierung	26
Begriff	27
Einteilung	28
Diagrammtyp	29
6.8 ZEITSCHRIFTEN - 3D-Streudiagramm	30
TELECOM - Dimensionalität	31
VISUALISIERUNGSWERKZEUG - Datenauswahl	32
VISUALISIERUNGSWERKZEUG - Methode	33
VISUALISIERUNGSWERKZEUG - Dimensionalität	34
VISUALISIERUNGSWERKZEUG - Diagrammtyp	35
🖱 Visualisierung mit SPSS Diamond (A 6.2)	36
Data Mining-Werkzeug	41
6.9 Methoden und Werkzeuge im Überblick	42
6.10 Werkzeugkriterien	43
Wenige Methoden und viele Werkzeuge	44
Statistik-Produkte	45
Ein typisches Paket zur konventionellen Statistik	46
🖱 BANK - Eine Data Mining-Spezifikation (A 6.1)	47
🖱 Ausschnitt aus dem Datenmodell (A 6.1)	49
Data Mining auf Tabellenblättern?	50
Spreadsheet-Lücke Grafik	51
Spreadsheet-Lücke Multidimensionalität	52
Spreadsheet-Lücken Metadaten verwalten	53

<u>Spreadsheet-Lücke Fehlende Werte</u>	<u>54</u>
<u>Spreadsheet-Lücke Eingabefehler beheben</u>	<u>55</u>
<u>Spreadsheet-Lücke Komfortables Pivoting</u>	<u>56</u>
<u>Spreadsheet-Lücke Grosse Datenmengen</u>	<u>57</u>
<u>Internet-Adressen</u>	<u>58</u>