Fachhochschule Heilbronn / Universität Heidelberg
Fachbereich Medizinische Informatik

Web Usage Mining
Eine Praxisanwendung im E-CRM

21. Februar 2003

Hussein Waly
Universität Heidelberg
Fachbereich Medizinische Informatik

Agenda

- Forschungsprojekt “Web Usage Mining”
  - Generelle Zielsetzung
  - Ausgangslage
- Web Mining
  - Definition
  - Web Mining-Prozess
  - Datenaufbereitung
- Web Mining in der Praxis
  - Beispiel bei einem Online-Shop
  - Logfile-Analyse
  - Beschreibung von Besucher-Profilen

Forschungsprojekt

- **Generelle Zielsetzung**
  - **Vorarbeit für ein Forschungsvorhaben an der FH Heilbronn**
  - **Einschätzung der Potenziale des „Web Usage Mining“ im E-CRM**
  - **Ermittlung welche Informationen in den Webserver-Logfiles zu finden sind**
  - **Welchen speziellen Mehrwert können diese Informationen für Data Mining-Fragestellungen im Marketing generieren?**

Forschungsprojekt

Ausgangslage

- **Netscape Webserver-Logfiles eines Online-Weinhandels in Karlsruhe**
- **Format: Extended Common Logfile Format „ECLF“**
- **Zeitraum: 4 Wochen**
- **Größe: insgesamt 110 Mbyte**
- **Analysetools:**
  - **SAS/WebHound™: Logfile-Analyse- und Reporting-Tool**
  - **SAS/Enterprise Miner™: Zum Einsatz von Data Mining-Verfahren**

Forschungsprojekt

- **Web Mining-Fragestellungen**

  Folgende Fragestellungen sind unter Anwendung verschiedener Data Mining-Verfahren zu bewältigen:

  - **Woher kommen die Besucher (Länder, Referrer-Seiten, Organisationen etc.)?**
  - **Welcher Browser wird verwendet?**
  - **Welche sind die Top Einstiegs-Seiten?**
  - **Lassen sich aus den Logfiles konkrete Besucher- bzw. Navigations-Profile ableiten?**
  - **Was sind die Einflussfaktoren auf einen Bestellvorgang des Katalogs im Online-Shop?**

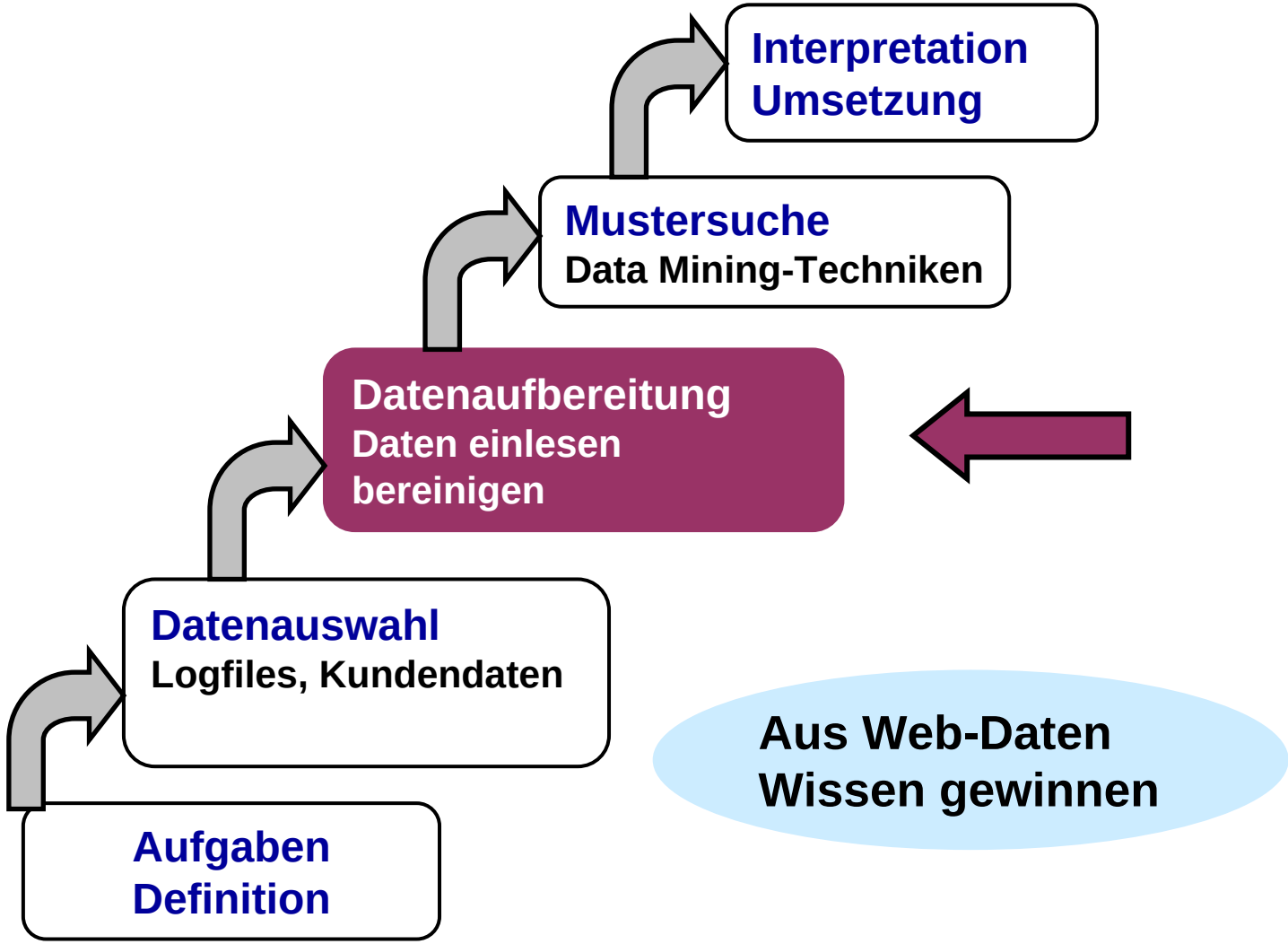
Web Mining-Begriffe

- **Definition**

  Anwendung von Data Mining-Verfahren auf Internet-Daten

Web Content Mining	Web Structure Mining	Web Usage Mining
Direkte Analyse der Seiten-Inhalte	Analysen der Linkstruktur einer Website	Analyse und Prognose des Besucher-Verhaltens
Einfache Erkennung und Gestaltung von Web-Dokumenten	Typisierung der Seiten (Einstiegs- ,Verteiler-, Inhaltsseiten)	Gängige Quellen sind : Logfiles und Einbindung von Zusatzdaten
Einsatz von Text Mining		

Web Mining-Prozess



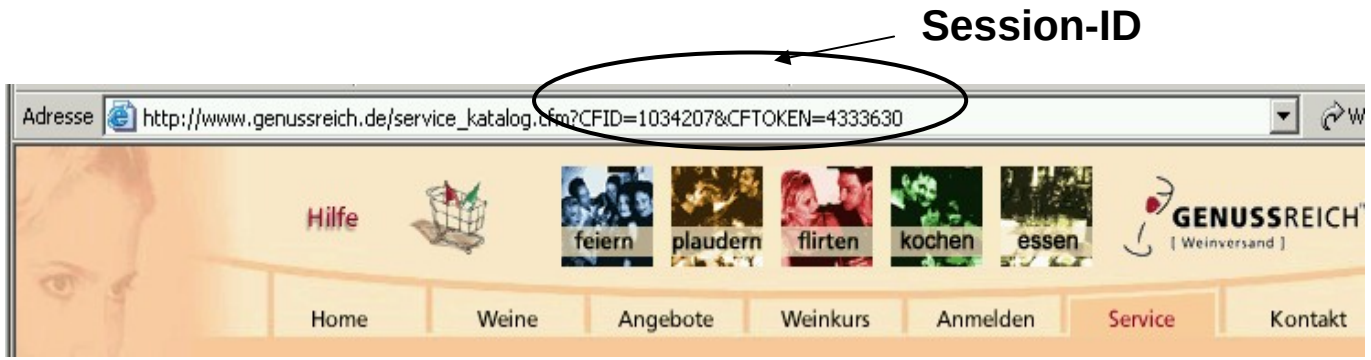
Datenaufbereitung

- **Technische Erweiterungen der Logfiles**
  - ➲ **werden operativ auf dem Webserver eingesetzt**

Beispiele:

– **Verwendung von Session-IDs**

Vergabe eindeutiger Kennung des Besuchers während einer Session (Sitzung)



Datenaufbereitung

- **Technische Erweiterungen der Logfiles**

  Beispiele:

  - **Verwendung von Cookies**
    - **Eindeutige Identifikation des Browsers eines Besuchers**
    - **Cookies werden auf Festplatte des Besuchers abgelegt (nicht in der URL)**
    - **Lösung der Problematik des Proxy-Servers und der dynamischen IP-Adresse**
  - **Registrierung der Besucher (User-IDs)**
    - **Dabei lassen sich alle Aktivitäten des Besuchers auf der Website nachvollziehen**

Einträge in einem Logfile

```
129.13.122.23 - - [20/Oct/1999:10:00:29 +0200] "GET /webmining/intern/preview/toc.xmlfrag HTTP/1.0" 200 216
129.13.122.23 - - [20/Oct/1999:10:00:29 +0200] "GET /webmining/intern/preview/toc.xmlfrag HTTP/1.0" 200 216
134.155.17.201 - - [20/Oct/1999:10:00:30 +0200] "GET /webmining/ HTTP/1.0" 200 5797
134.155.17.201 - - [20/Oct/1999:10:00:32 +0200] "GET /webmining/slide.css HTTP/1.0" 304 -
134.155.17.201 - - [20/Oct/1999:10:00:33 +0200] "GET /icons/webmining/tocright.gif HTTP/1.0" 304 -
134.155.17.201 - - [20/Oct/1999:10:00:33 +0200] "GET /icons/greenball.gif HTTP/1.0" 304 -
134.155.17.201 - - [20/Oct/1999:10:00:33 +0200] "GET /icons/etufo.gif HTTP/1.0" 304 -
129.13.122.23 - - [20/Oct/1999:10:00:36 +0200] "GET /webmining/intern/preview/toc.xmlfrag HTTP/1.0" 200 216
129.13.122.23 - - [20/Oct/1999:10:00:36 +0200] "GET /webmining/intern/preview/toc.xmlfrag HTTP/1.0" 200 216
134.155.17.201 - - [20/Oct/1999:10:00:37 +0200] "GET /webmining/Script-1.xml HTTP/1.0" 200 2813
134.155.17.201 - - [20/Oct/1999:10:00:38 +0200] "GET /icons/webmining/tocleft.gif HTTP/1.0" 304 -
129.13.122.23 - - [20/Oct/1999:10:01:00 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
129.13.122.23 - - [20/Oct/1999:10:01:00 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
134.155.17.201 - - [20/Oct/1999:10:01:01 +0200] "GET /webmining/script/1/ HTTP/1.0" 200 4379
134.155.17.201 - - [20/Oct/1999:10:01:02 +0200] "GET /webmining/script/1/slide.css HTTP/1.0" 304 -
129.13.122.23 - - [20/Oct/1999:10:01:39 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
129.13.122.23 - - [20/Oct/1999:10:01:39 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
134.155.17.201 - - [20/Oct/1999:10:01:41 +0200] "GET /webmining/script/1/ HTTP/1.0" 200 4379
134.155.17.201 - - [20/Oct/1999:10:01:42 +0200] "GET /robots.txt HTTP/1.0" 200 164
134.155.17.201 - - [20/Oct/1999:10:01:46 +0200] "GET /webmining/script/1/titlepage-2.xml HTTP/1.0" 200 1676
134.155.17.201 - - [20/Oct/1999:10:01:51 +0200] "GET /icons/webmining/tocright.gif HTTP/1.0" 200 172
134.155.17.201 - - [20/Oct/1999:10:01:52 +0200] "GET /icons/greenball.gif HTTP/1.0" 200 398
134.155.17.201 - - [20/Oct/1999:10:01:53 +0200] "GET /icons/etufo.gif HTTP/1.0" 200 1490
129.13.122.23 - - [20/Oct/1999:10:01:55 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
129.13.122.23 - - [20/Oct/1999:10:01:55 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
134.155.17.201 - - [20/Oct/1999:10:01:56 +0200] "GET /webmining/script/1/OrgI-1.xml HTTP/1.0" 200 3729
129.13.122.23 - - [20/Oct/1999:10:01:58 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
129.13.122.23 - - [20/Oct/1999:10:01:58 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
134.155.17.201 - - [20/Oct/1999:10:01:58 +0200] "GET /webmining/script/1/OrgI-2.xml HTTP/1.0" 200 3269
129.13.122.23 - - [20/Oct/1999:10:02:00 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
129.13.122.23 - - [20/Oct/1999:10:02:00 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
134.155.17.201 - - [20/Oct/1999:10:02:01 +0200] "GET /webmining/script/1/OrgI-3.xml HTTP/1.0" 200 4268
134.155.17.201 - - [20/Oct/1999:10:02:11 +0200] "GET /webmining/script/1/OrgI-4.xml HTTP/1.0" 200 1671
134.155.17.201 - - [20/Oct/1999:10:02:12 +0200] "GET /icons/webmining/tocleft.gif HTTP/1.0" 200 172
129.13.122.23 - - [20/Oct/1999:10:02:13 +0200] "GET /webmining/intern/preview/script/1/toc.xmlfrag HTTP/1.0" 200 363
```

Datenaufbereitung

- **Schritte zur Datenaufbereitung**
  - **werden auf die bereits angefallenen Logfiles angewendet**

1) Data Cleaning

2) Benutzer- und Session-Identifikation

3) Pfadvervollständigung

Datenaufbereitungs-Schritte

Art der Datenaufbereitung	Aufgabe
Data Cleaning	• Eliminierung irrelevanter Logfile-Einträge z.B. automatischer Aufruf von Bild-Dateien u. Spider-Zugriffen müssen identifiziert und entfernt werden • Nur Benutzer-Aktionen sind von Interesse
Benutzer-Identifikation und Session-Identifikation	• Zuordnung von Logfile-Einträgen zu einzelnen Benutzern • IP-Adressen identifizieren Benutzer nicht eindeutig z.B. Proxy-Sever u. dyn. IPs • Gliederung der Zugriffe in Sessions
Pfadvervollständigung	• Ergänzung fehlender Zugriffe in einem Zugriffspfad • Bookmarks u. Anfragen aus dem Cache

Web Mining-Verfahren

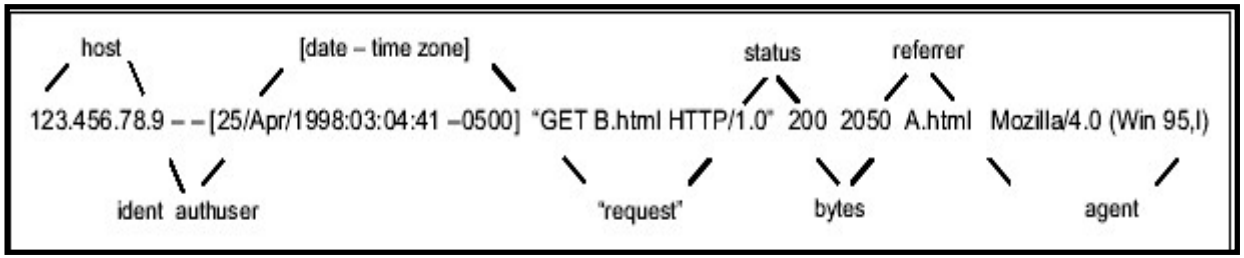
Aufgabenstellung	Web Mining-Verfahren
Analyse von Navigationspfaden **Welcher Navigationspfad führt zu einer Katalogbestellung?**	• **Assoziationsanalyse** • **Sequenzanalyse**
Erkennung von Besuchertypen **Welche Besuchergruppe bestellt den Katalog?**	• **Clusteranalyse** • **Kohonen SOM**
Vorhersage / Beschreiben von Besucherverhalten **Was unterscheidet einen Besucher von einem Katalogbesteller?**	• **Entscheidungsbaum** • **Regressionsanalyse** • **Neuronale Netze**

Web Mining in der Praxis

- **Praxisbeispiel mit dem SAS/WebHound™ und SAS/Enterprise Miner™**
  - **Logfile-Analyse**
  - **Beschreibung von Besucher-Profilen**
- **Als Datenquelle dienen die Logfiles des Webservers eines Online-Shops für Weinprodukte (www.genussreich.de)**
- **Netscape Webserver im „Extended Common Log Format“**

Praxisdatensatz

Feldname	Bedeutung
Host	**IP-Adresse**
Date	**Datum und Uhrzeit**
Timezone	**Abweichung von GMT in Stunden**
Request	**Methode und Dokument**
Status	**Codenummer (200:OK)**
Bytes	**Gesamtzahl der übertragenen Bytes**
Referrer	**URL der Seite, die den Link zur angefragten Seite enthält**
Agent	**Browser (Typ u. Version)**

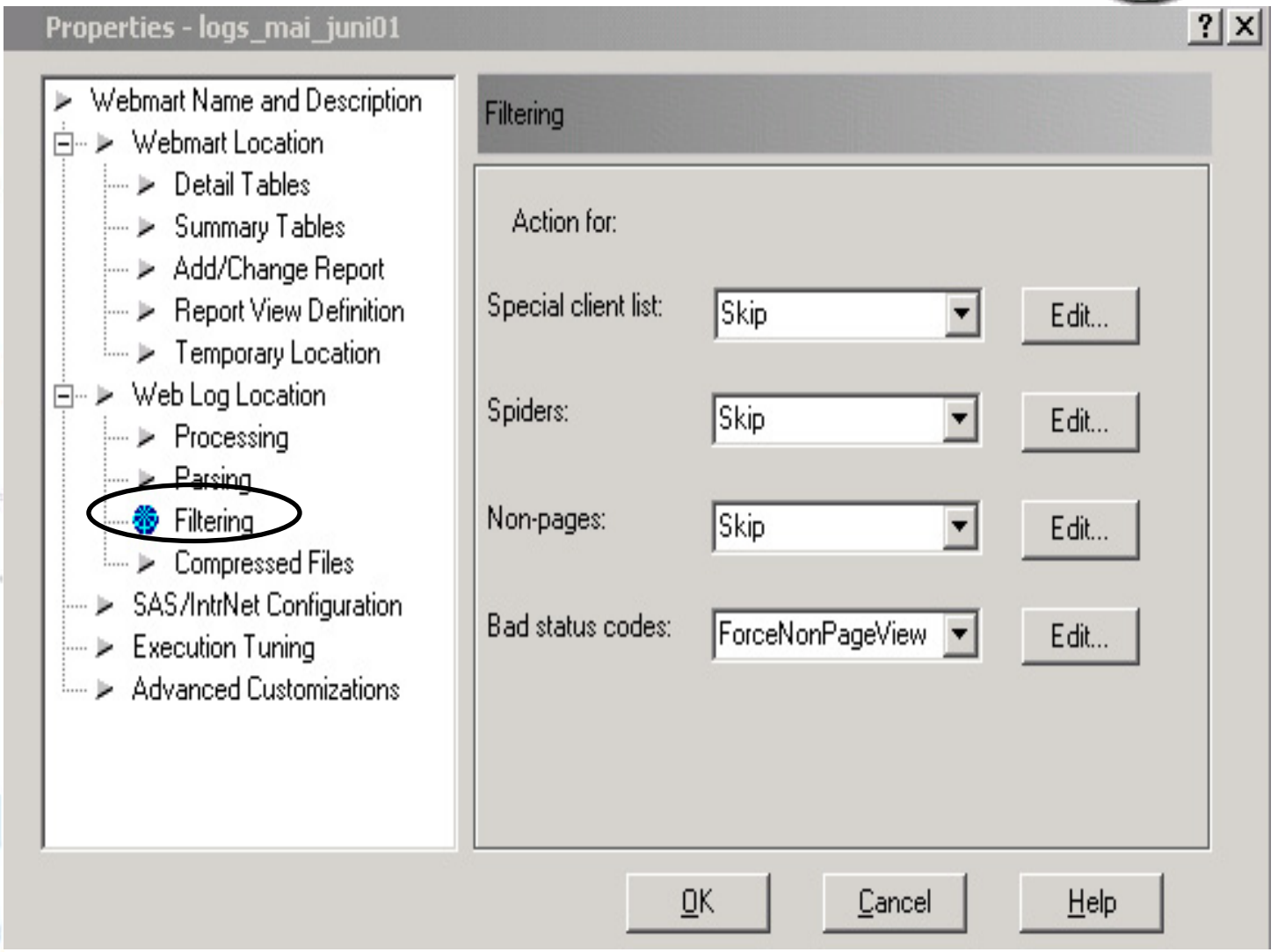


SAS/WebHound™

- **Logfile-Anlayse- und Web-Reporting-Tool**
- **besteht aus: SAS Base 8.2, SAS Graph, SAS IntrNet, SAS AF, SAS Connect, SAS ETS , SAS OLAP Server und SAS Warehouse**
- **Logfiles einlesen und aufbereiten unter Einbindung externer Daten**
- **Verdichtung der Daten, Erstellung von SAS Data Sets und MDDBs für OLAP-Reporting**
- **ca. 300 Standard Reports, Explorerartiger Report Viewer**

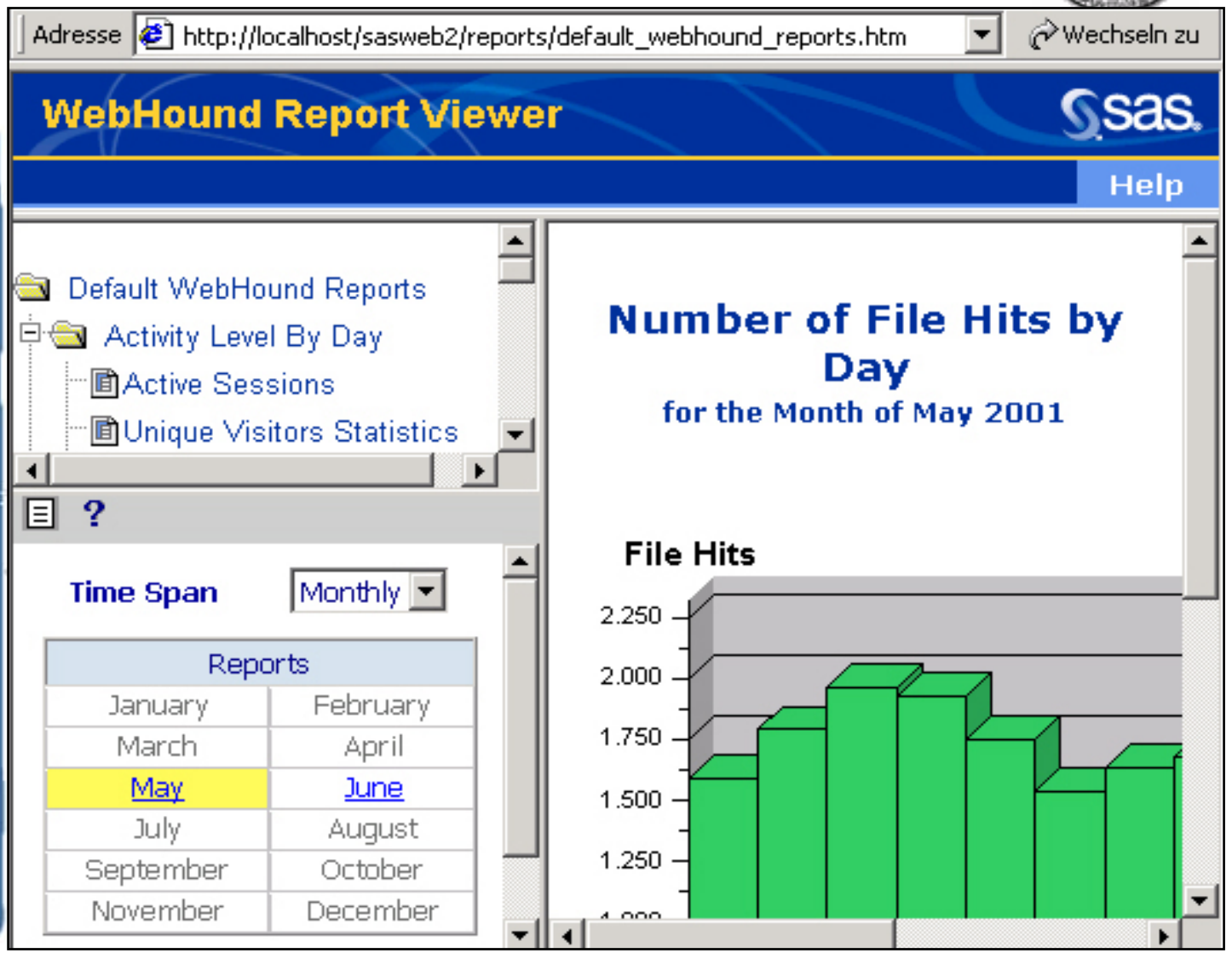
Datenaufbereitung

- Zielsetzung
- Ausgangslage
- Web Mining
- Praxisbeispiel
- **SAS WebHound™**
- Fazit

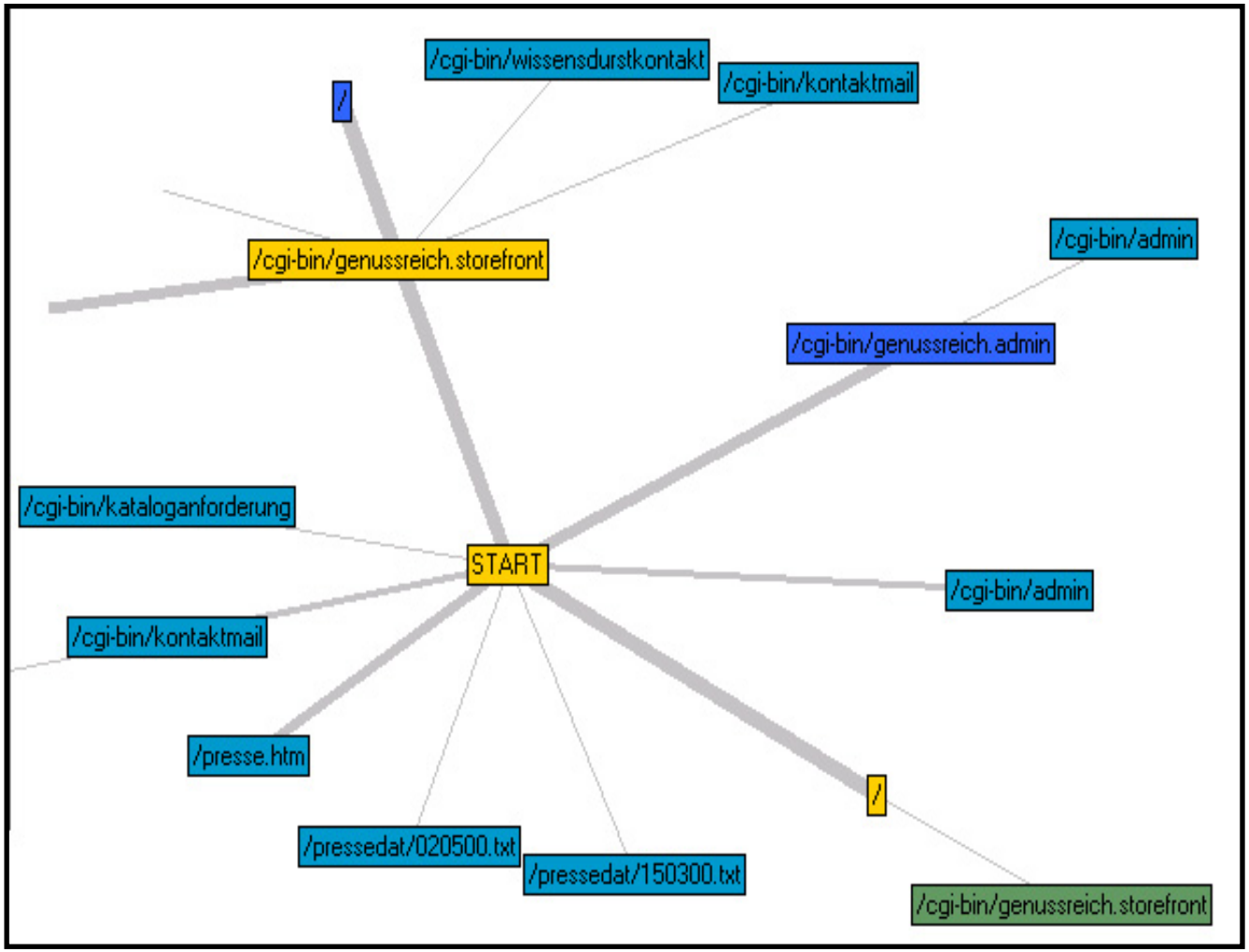


Report Viewer

- Zielsetzung
- Ausgangslage
- Web Mining
- Praxisbeispiel
- **SAS WebHound™**
- Fazit



Surfverhalten der Besucher



Woher kommen die Besucher?

Germany	692
Austria	62
Switzerland	20
Netherlands	5
Canada	4
China	2
France	2
Tonga	2
Russia	1
Sweden	1
Ukraine	1
Total	5,012

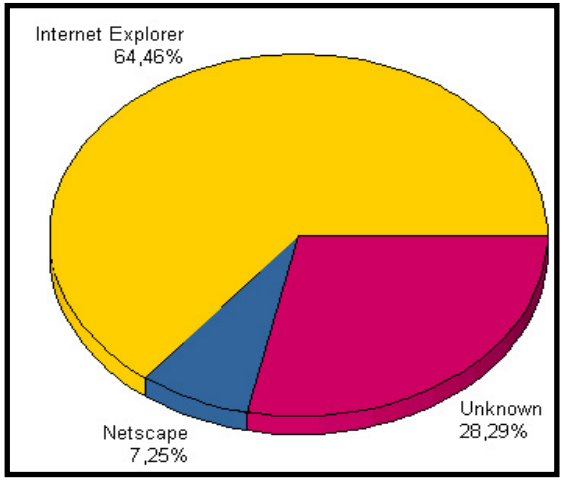
2	banner-srv.fairad.de	1,202
3	www.genussreich.de	977
4	www.aol.de	199
5	www.payback.de	39
6	shopping.msn.de	15
7	member.payback.de	13
8	www.paybox.de	13
9	www.wein-plus.de	9
10	www.google.de	7

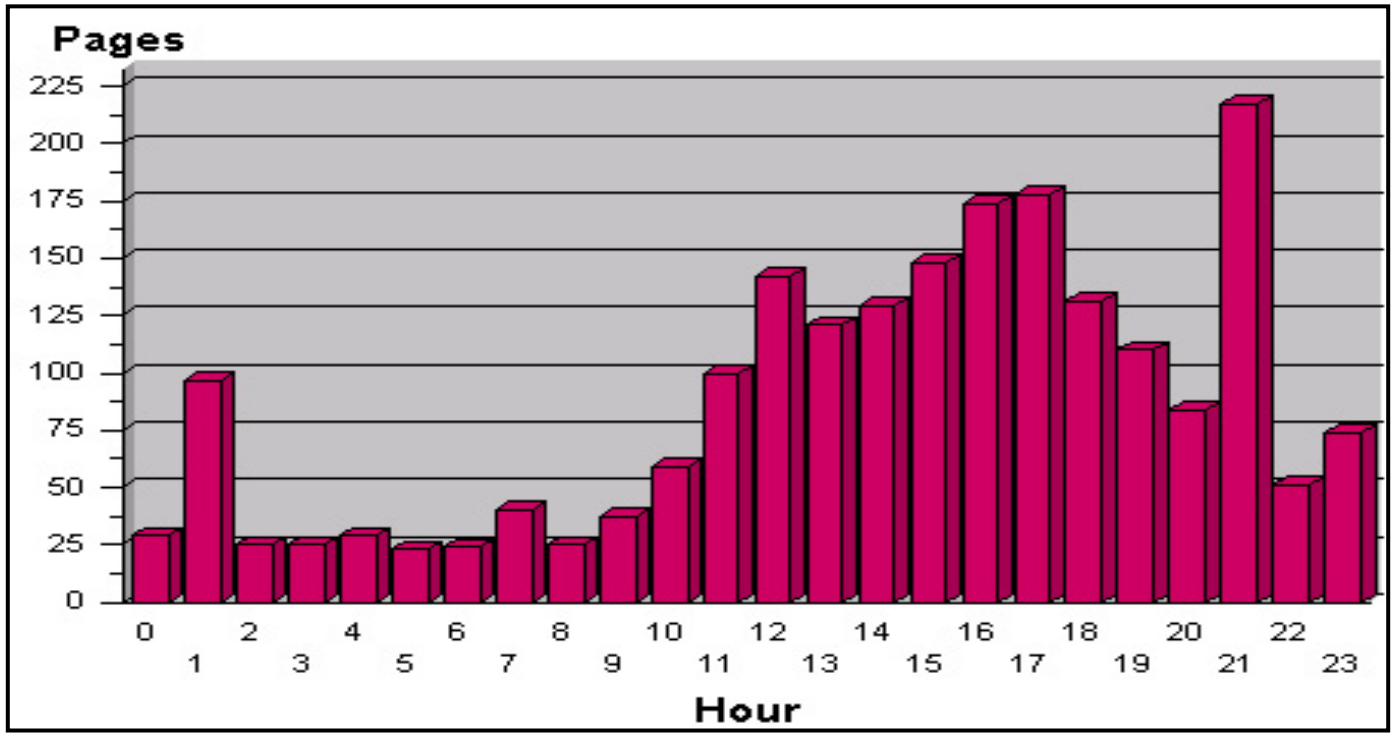
Top 10 Einstiegs-Seiten

- Zielsetzung
- Ausgangslage
- Web Mining
- **Praxisbeispiel**
- Treeview
- Referreranalyse
- **Seitenanalyse**
- Fazit

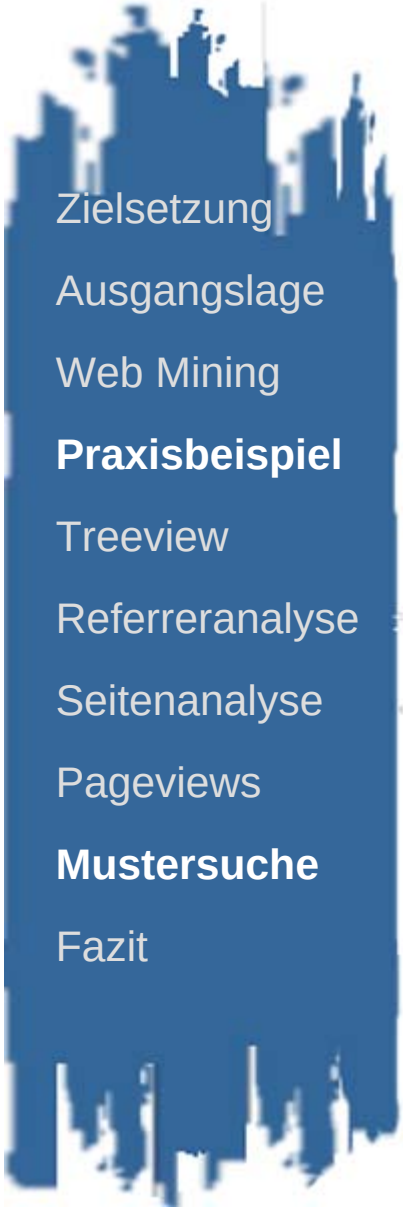
Top Ten Entry Points and Percent of Total			
Rank	Requested File	Entry Point Count	Percent of Total
1	/cgi-bin/genussreich.storefront	3,373	67%
2	/	1,570	31%
3	/cgi-bin/genussreich.admin	55	1.1%
4	/cgi-bin/admin	5	.10%
5	/cgi-bin/kontaktmail	2	.04%
6	/presse.htm	2	.04%
7	/pressedat/genussreich.txt	2	.04%
8	/cgi-bin/genusstest.storefront	1	.02%
9	/cgi-bin/kataloganforderung	1	.02%
10	/pressedat/150300.txt	1	.02%
		5,012	*100%*

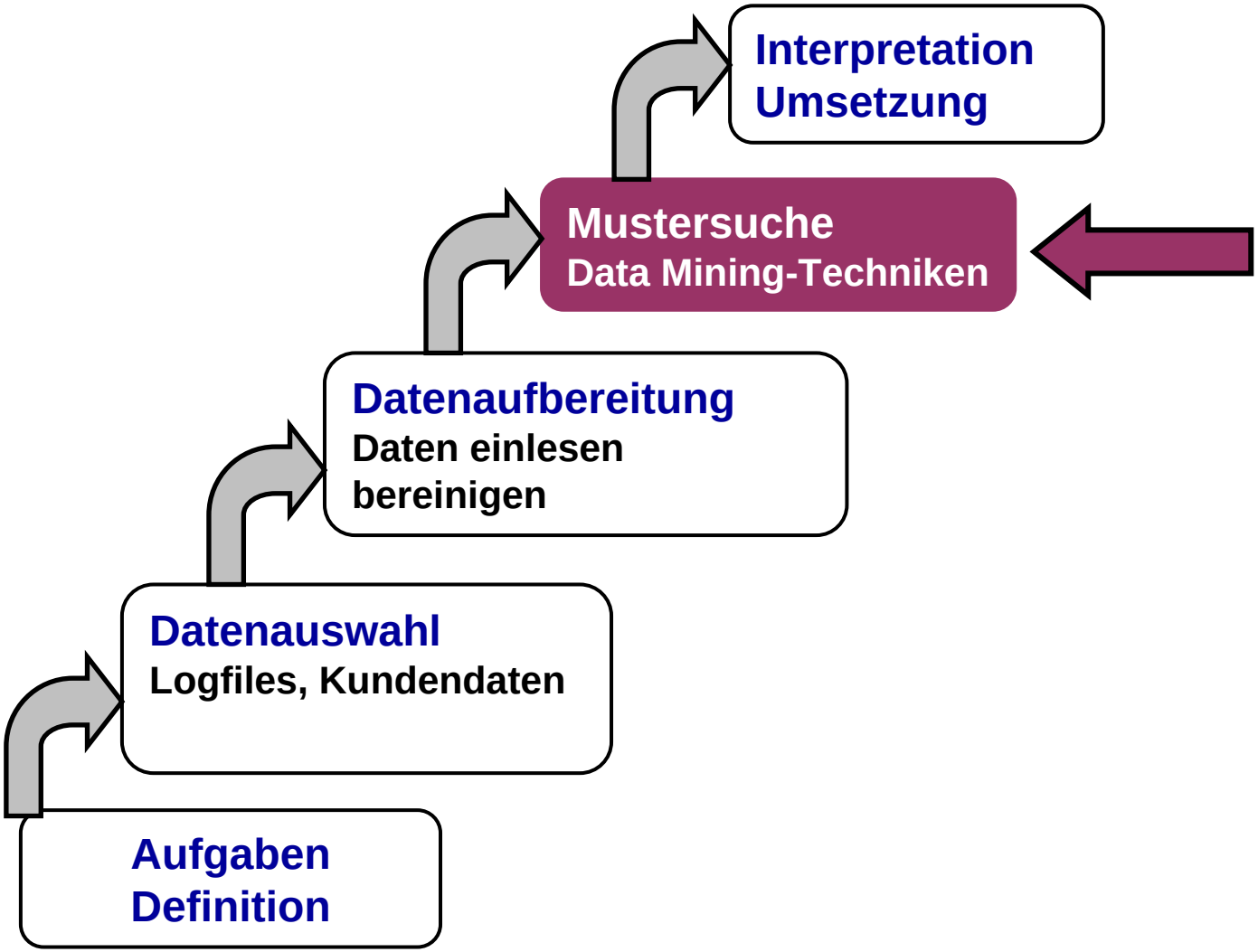
Browser und Pageviews





Web Mining in der Praxis





Beschreibung von Besucher-Profilen

- **Was sind die Einflussfaktoren auf einen Bestellvorgang des Katalogs im Online-Shop?**
  - **Einsatz von Data Mining-Verfahren mit dem SAS/Enterprise Miner™**
- **Anwendung von Web Mining-Segmentierungsmodellen (z.B. Entscheidungsbaum-Verfahren)**
- **Weitere Datenaufbereitungs-Schritte werden benötig (Transformation und 0/1- Kodierung)**

Vorgehensweise der Analyse

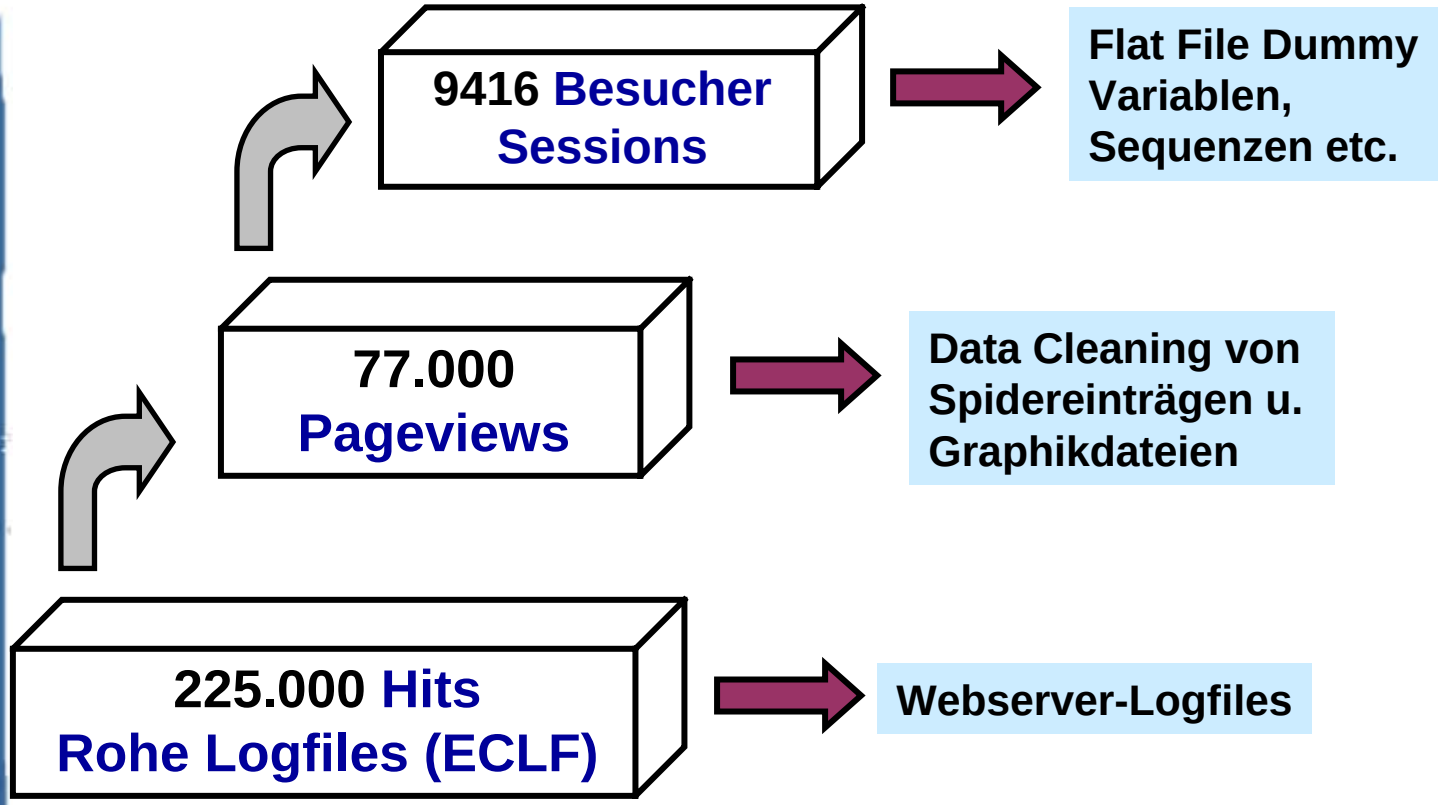
- **Einlesen, Aufbereiten der Logfiles und Erstellung von SAS Data Sets (mit 77.000 Datensätzen)**
- **Folgende Informationen liegen vor:**
  - **Session-IDs (IP, Zeitstempel, User Agent, BS)**
  - **aufgerufene Webseiten**
  - **Referrer-URL**
  - **Dauer und Startzeit einer Session**
  - **Katalog bestellt (Ja / Nein)**
- **Erstellung eines „Flat Files“ durch Tabellentransformation, Sequenzbildung und Einführung von Dummy-Variablen**

Vorgehensweise der Analyse

- **Sequenzbildung:**
  Feststellung der häufigsten Pfade mit der Sequenzanalyse
- **Anschließend erfolgt eine 0/1-Kodierung der Sequenzen mit hohem Support u. Confidence**
- **Insgesamt reduziert sich die Anzahl der Datensätze von 77.000 auf 9416 Datensätze im „Flat File“**
- **Flat File --> Datenbasis des Data Mining-Modells**

Weg zu Trainingsdaten

- Zielsetzung
- Ausgangslage
- Web Mining
- **Praxisbeispiel**
- Treeview
- Referreranalyse
- Seitenanalyse
- Pageviews
- **User-Profil**
- Fazit



Trainingsdaten

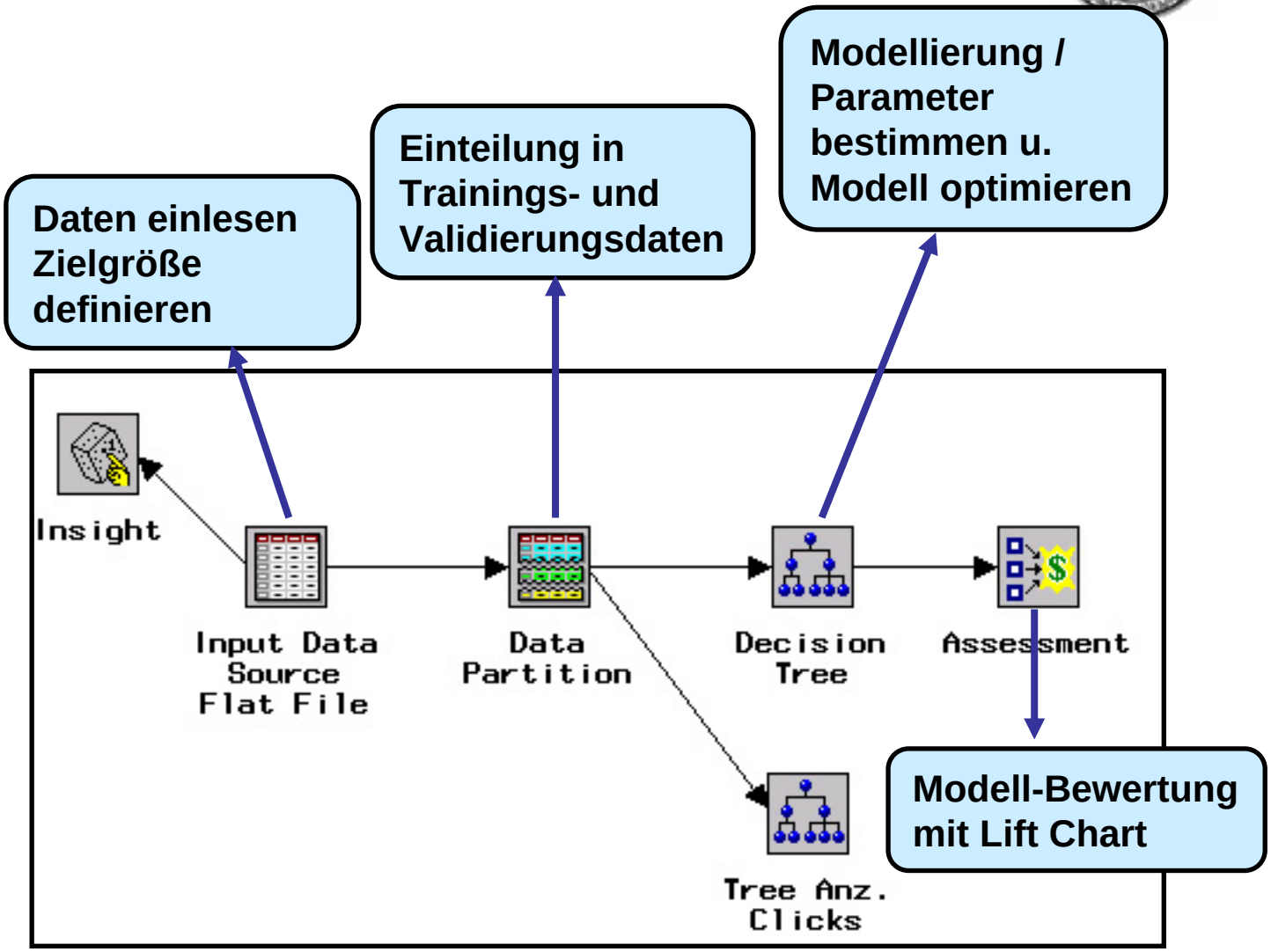
• Zielvariable über Katalogbestellung (0/1) kodiert	• Datum einer Session
• Session-Identifikation	• Start einer Session
• IP-Adresse des Besuchers	• Dauer einer Session
• Referrer-Seite des Besuchers	• Anzahl der Klicks
• Unterscheidung Werktag / Wochentag	• 10 Sequenzvariablen
• Verweildauer pro Webseite	• 32 Variablen für die Webseiten

Data Mining nach der SEMMA-Methodik von SAS

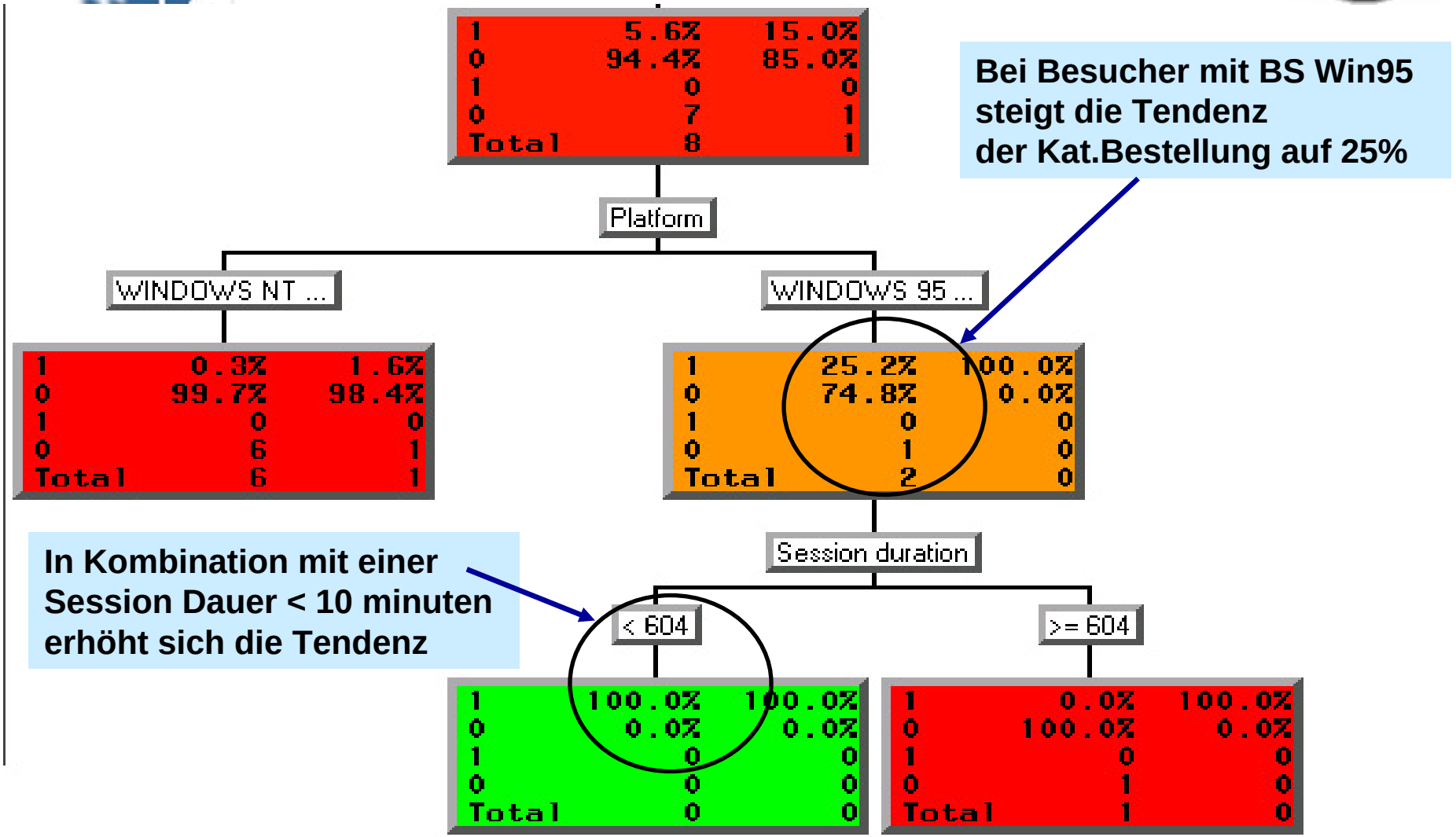
Sample, Explore, Modify, Model, Assess

Modellierung

- Zielsetzung
- Ausgangslage
- Web Mining
- **Praxisbeispiel**
- Treeview
- Referreranalyse
- Seitenanalyse
- Pageviews
- **Modell**
- Fazit



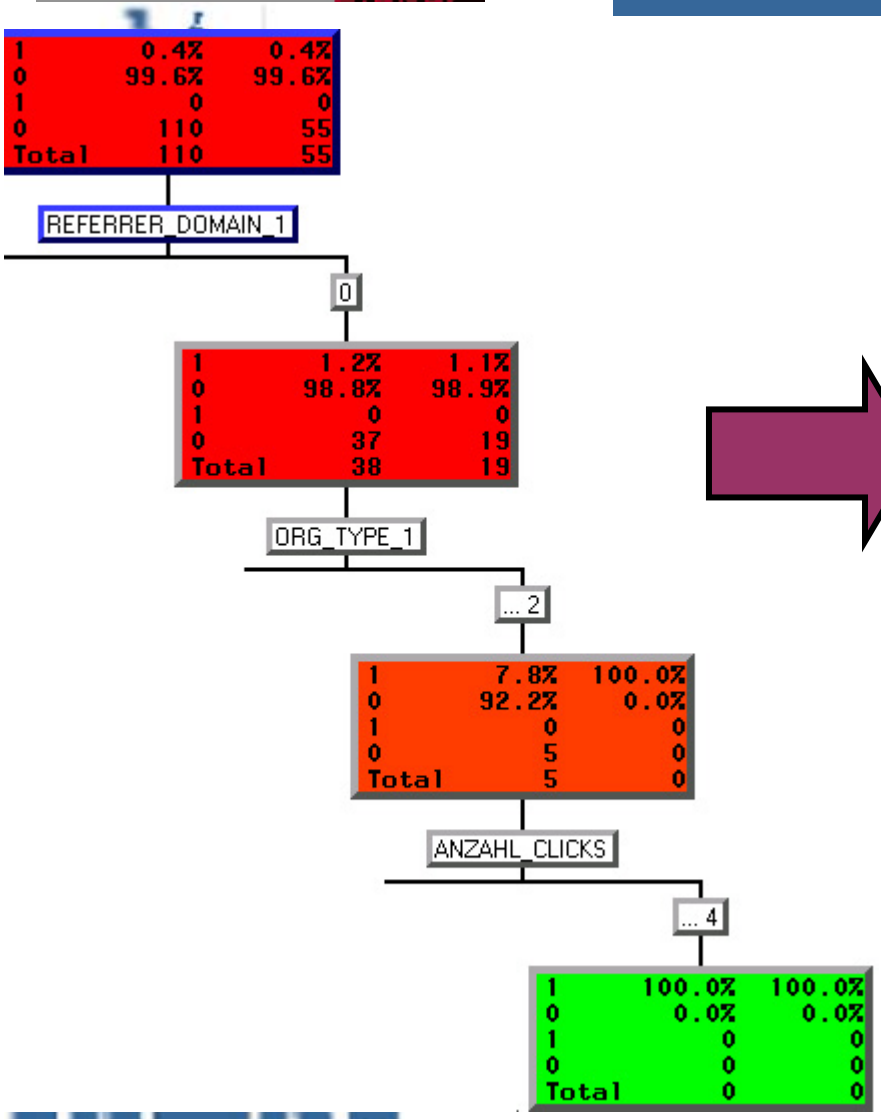
Entscheidungsbaum User-Profil (1)



Bei Besucher mit BS Win95 steigt die Tendenz der Kat.Bestellung auf 25%

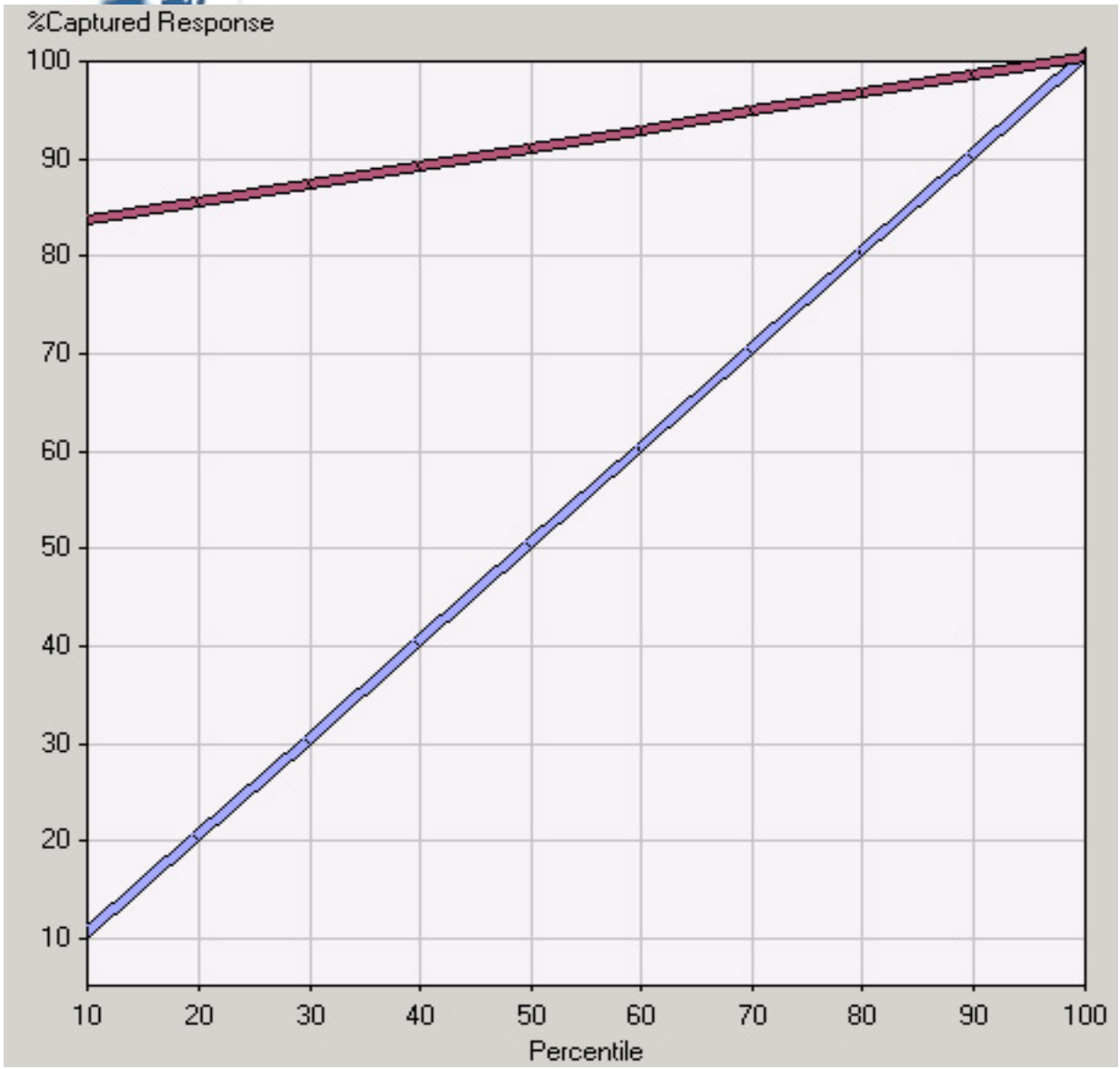
In Kombination mit einer Session Dauer < 10 minuten erhöht sich die Tendenz

Entscheidungsbaum User-Profil (2)



Wenn ein Besucher von einer bestimmten Referrer-Seite (.com) kommt und tätigt bis 4 Clicks, dann erhöht sich die Bestell-Wahrscheinlichkeit deutlich

Modell-Bewertung
Lift Chart Captured Response



Bewertung der Güte des Verfahrens

Beispiel: bei der Auswahl der besten 20% werden bereits 85% der Katalogbesteller erfasst

H.Waly

Was bringt Web Usage Mining?

Das Besucherverhalten im Online-Shop zu verstehen

- **Somit kann beispielsweise der Workflow eines Bestellvorgangs optimiert werden (Erhöhung der Konversionsrate, höherer Umsatz , verbesserte Kundenbindung u. Kundenloyalität)**
- **Optimierung des Web-Auftritts (Angebot u. Seiten)**
- **Messung und Erhöhung der Effizienz von Bannerschaltungen**
- **Enterprise Miner ist eine wichtige Ergänzung für die Extraktion signifikanter Benutzer-Profile aus den Logfiles**

Vielen Dank für Ihre Aufmerksamkeit !

Hussein Waly
Universität Heidelberg
Fachbereich Medizinische Informatik
hwaly@gmx.de