Universität Leipzig

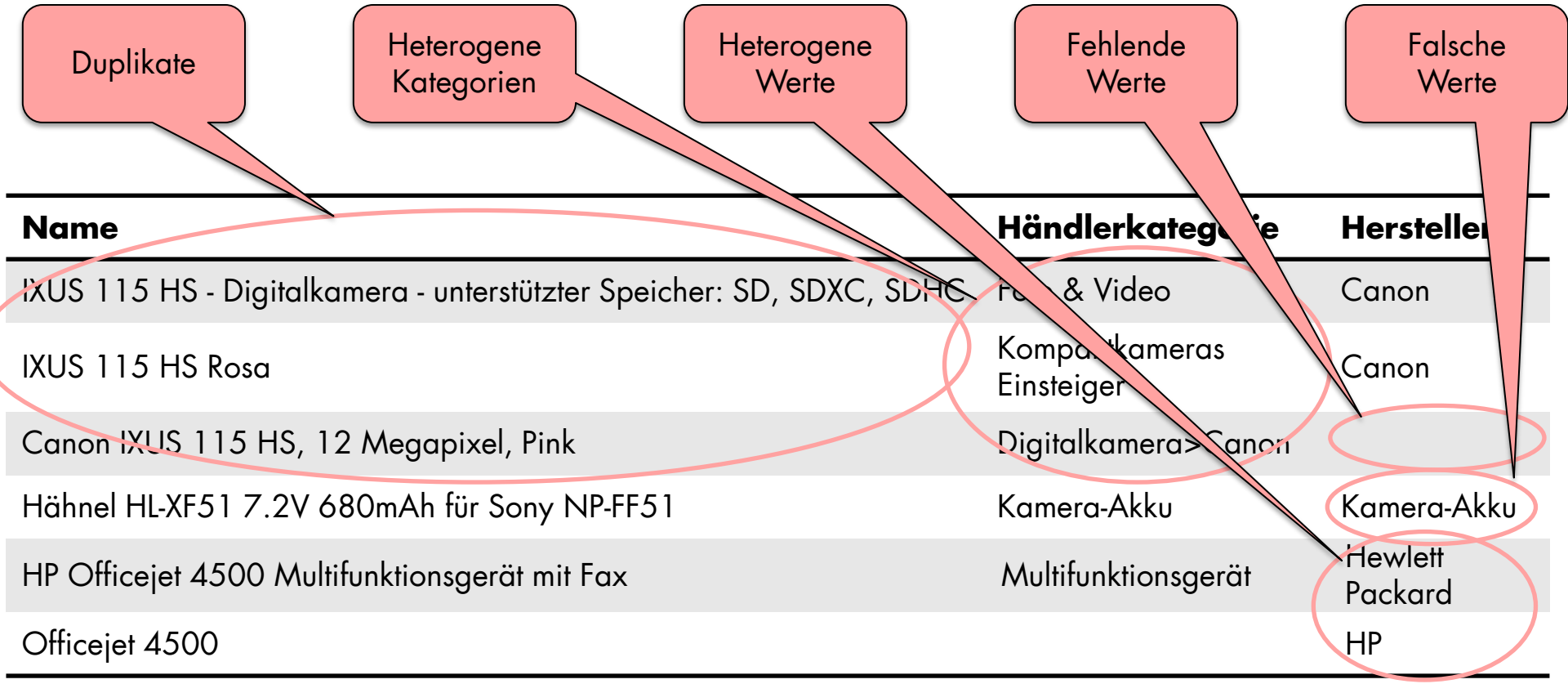
Institut für Informatik

Methoden zur Kategorisierung und Duplikaterkennung von Produktdaten

Hanna Köpcke

Web Data Integration Workshop 2011

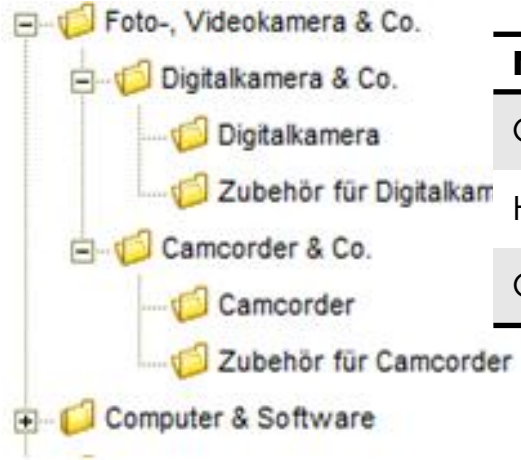
Was sind die Probleme?



Name	Händlerkategorie	Hersteller
IXUS 115 HS - Digitalkamera - unterstützter Speicher: SD, SDXC, SDHC	Foto & Video	Canon
IXUS 115 HS Rosa	Kompaktkameras Einsteiger	Canon
Canon IXUS 115 HS, 12 Megapixel, Pink	Digitalkamera>Canon	
Hähnel HL-XF51 7.2V 680mAh für Sony NP-FF51	Kamera-Akku	Kamera-Akku
HP Officejet 4500 Multifunktionsgerät mit Fax	Multifunktionsgerät	Hewlett Packard
Officejet 4500		HP

Lösungen

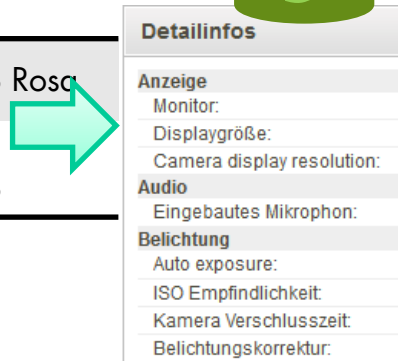
Kategorisierung | **Datenbereinigung** | **Duplikaterkennung**



Name	Hersteller	Code
Canon IXUS 115 HS	**Canon**	**IXUS 115 HS**
Hähnel HL-XF51 7.2V	**Hähnel** Akku	**HL-XF51**
Officejet 4500	**Hewlett Packard**	

IXUS 115 HS Rosa

IXUS 115 HS



	Kategorie
IXUS 115 HS Rosa	**Digitalkamera**
IXUS 105	**Digitalkamera**
Hähnel HL-XF51 7.2V	**Zubehör für Digitalkameras**

IXUS 115 HS Rosa

IXUS 115 HS

IXUS 105

IXUS 105

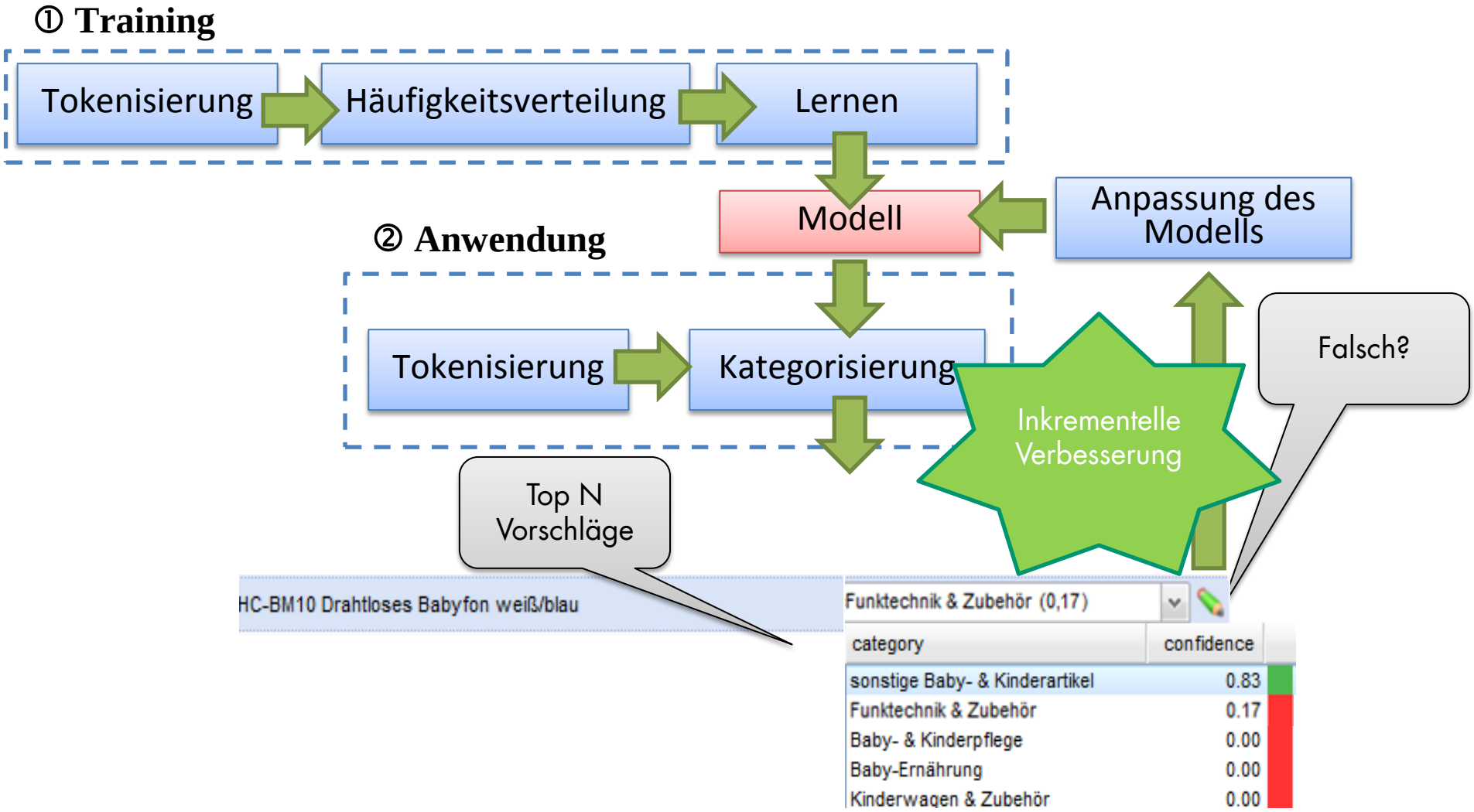
Kategorisierung

- Herausforderungen:
  - Große Anzahl an Kategorien und Angeboten
  - Große Unterschiede zwischen Katalogen
    - Struktur, Bezeichnungen
  - „Falsche“ Zuordnungen
    - Besseres Placement (Zubehör in Hauptkategorie)
  - Mehrere mögliche Kategorisierungen

Identifier:	14978321
Name:	HC-BM10 Drahtloses Babyfon weiß/blau
Beschreibug:	Alarm beim Verlassen des Empfangsbereichs LED
Hersteller:	König
Preis:	65.95
Bild:	

Babyartikel?
Funktechnik?

(Semi-)automatische Kategorisierung



Datenbereinigung

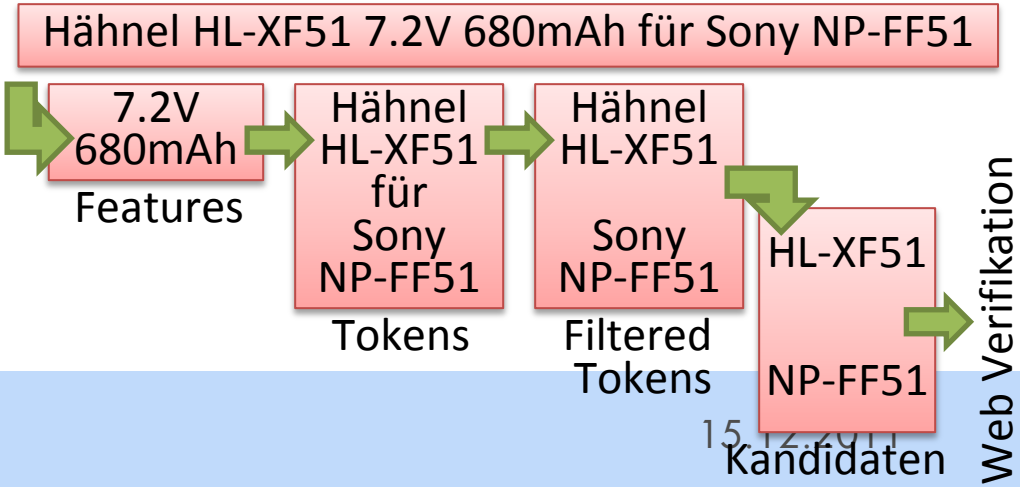
- Cleaning kritischer Attribute, z.B. von Hersteller- oder Markenangaben
  - Automatisierte Erstellung von Lookup-Tabellen mit Synonym-Verwaltung

Allied Telesis	Allied Telesis
Allied	Allied Telesis
Allied Telesyn	Allied Telesis

- Feature-Extraktion und -Verbesserung
  - Produkt-Ids, z.B. EAN, GTIN

HC-BM10 Drahtloses Babyfon weiß/blau
Alarm beim Verlassen des Empfangsbereichs LED EAN: 5412810142194

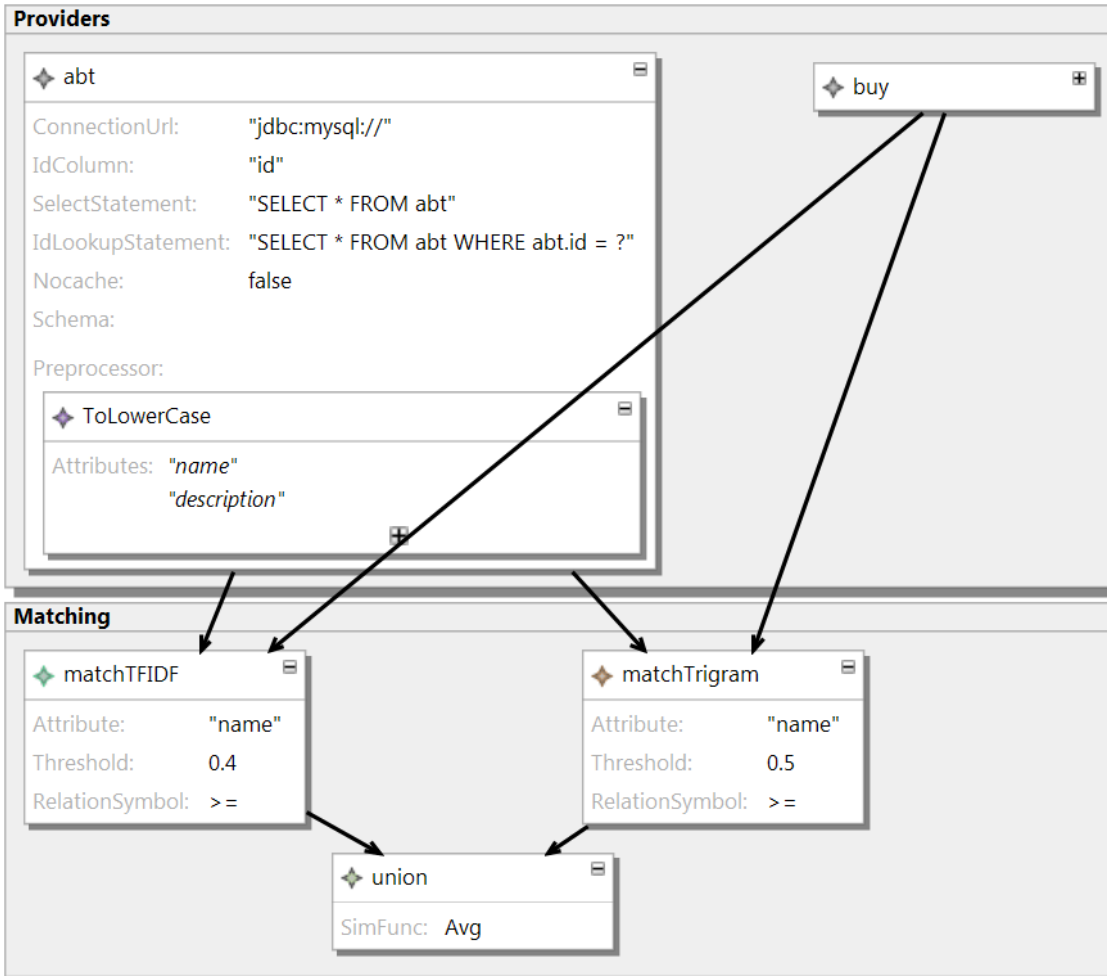
  - Produkt-Codes



Matchen zur Erkennung von Duplikaten

- Matchen = Identifikation semantisch äquivalenter Objekte
- Matchen von Produktangeboten
  - zu Produkten in einem Produktkatalog
  - mit anderen Angeboten
- Flexible Kombination mehrerer Match-Verfahren im Rahmen von Objekt-Matching-Workflows

Match-Workflow

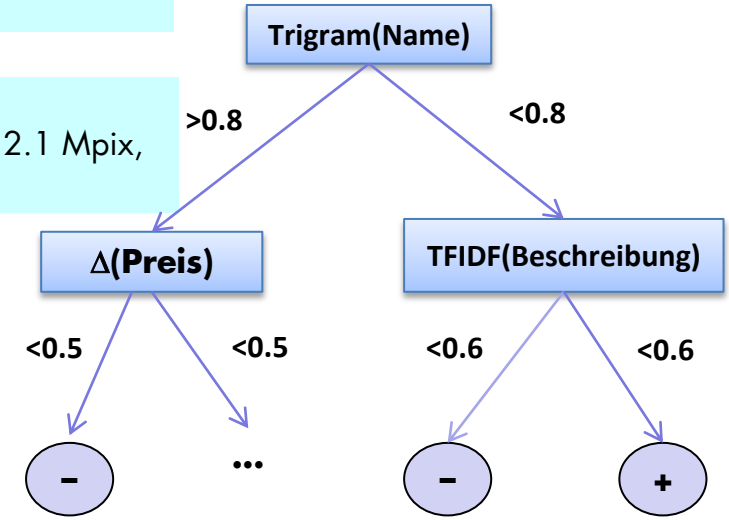


Trainingsbasierte Match-Verfahren

- Reduzierung des manuellen Tuning-Aufwandes
- Trainingsdaten
  - Unterstützung bei der manuellen Annotation

IXUS 115 HS Rosa Digitale Kompaktkamera mit 12,1 Megapixel	=	**IXUS 115 HS** Digitalkamera, Kompaktkamera, 12.1 Mpix, 4-fach optischer Zoom
IXUS 115 HS Digitalkamera, Kompaktkamera, 12.1 Mpix, 4-fach optischer Zoom	≠	**IXUS 105** Digitalkamera, Kompaktkamera, 12.1 Mpix, 4-fach optischer Zoom

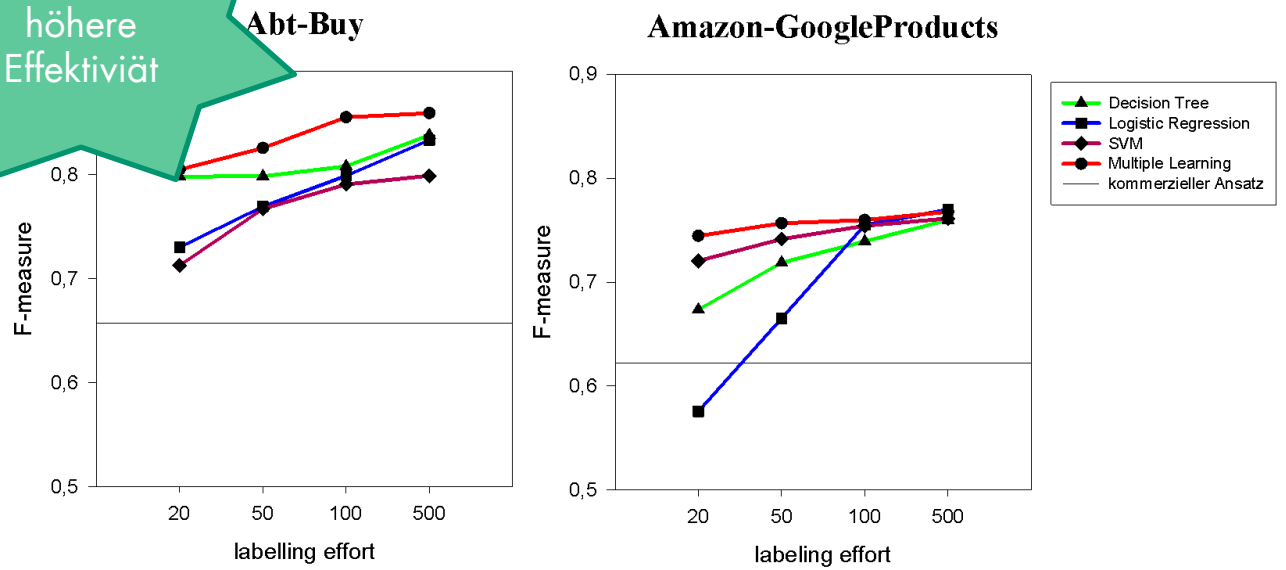
- Maschinelle Lernverfahren:
  - Entscheidungsbaum, SVM, ANN
    Logistische Regression, Mehrheitslerner

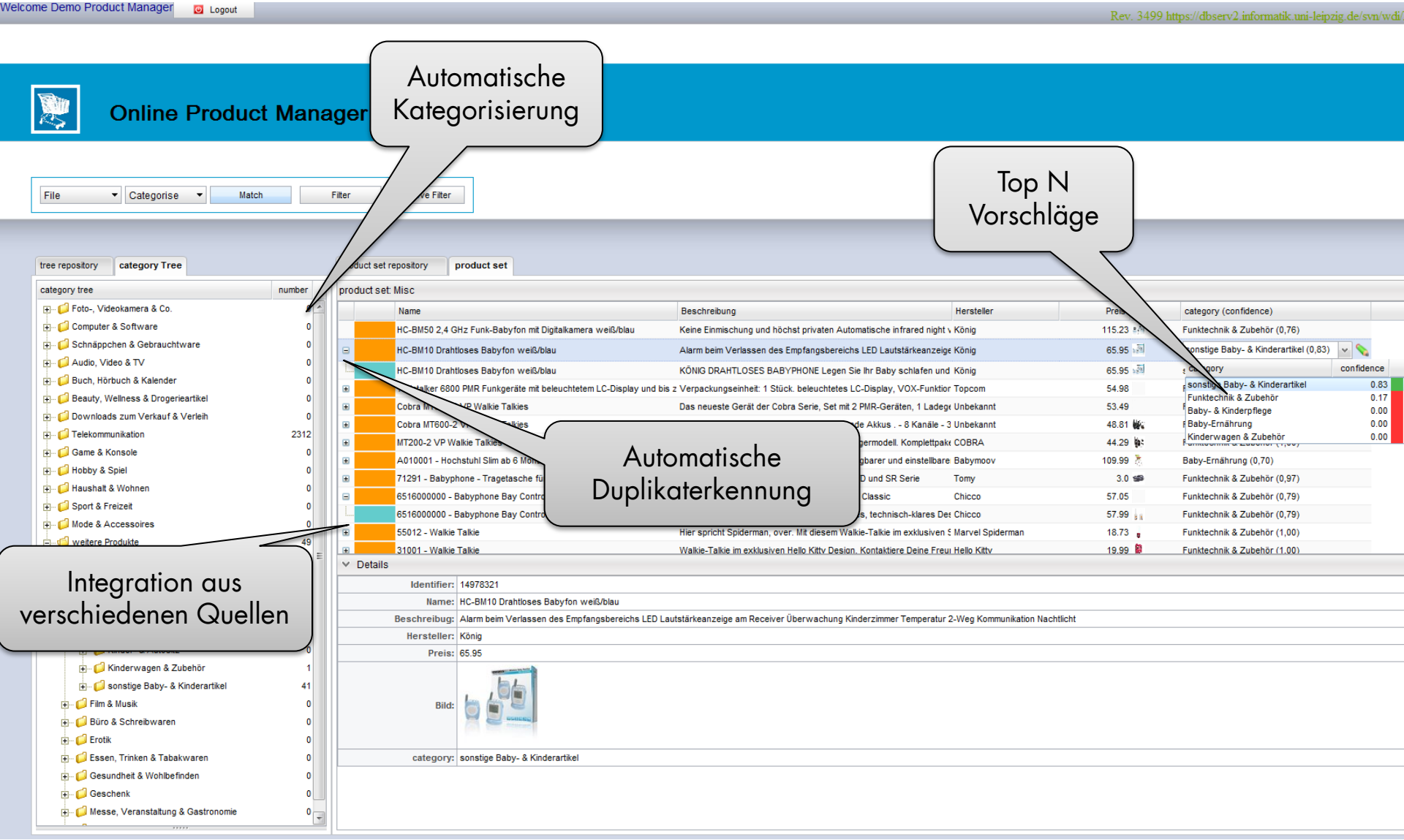


Vorteile unserer Lösungen

- Erhöhung der Datenqualität
- Einsparung von Zeit und Kosten
- Optimiert für große Datenmengen
- Branchenübergreifend einsetzbar

Bis zu 15 % höhere Effektiviät




Welcome Demo Product Manager
Logout
Rev. 3499 https://dbserv2.informatik.uni-leipzig.de/svn/wdi/
Online Product Manager
Automatische Kategorisierung
Top N Vorschläge
Automatische Duplikaterkennung
Integration aus verschiedenen Quellen
File
Categorise
Match
Filter
tree repository
category Tree
product set repository
product set
category tree
number
product set: Misc
Name
Beschreibung
Hersteller
category (confidence)
Details
Identifier:
Name:
Beschreibug:
Hersteller:
Preis:
Bild:
category:

Ausblick

- 2 Varianten verfügbar im **März 2012**
- **Community Edition**
  - Open Source, kostenlos
  - Einfache Matchstrategien

- **Software Suite** mit voller Funktionalität für Kategorisierung und Matching
  - für Business-Anwendungen