

Schlechte Daten – falsche Analyse

Carsten Kraus, Omikron GmbH
Pforzheim

Datenbanken sind wie Kaffeetassen



- Auf den Inhalt kommt es an!

5 Arten Schlechte Daten

- Unstrukturierte Daten
- Fehlende Daten
- Falsche Daten
- Veraltete Daten
- Dubletten

5 Arten Schlechte Daten

- Unstrukturierte Daten
- Fehlende Daten
- Falsche Daten
- Veraltete Daten
- Dubletten

Unstrukturierte Inhalte

- In 33000 Datensätzen: über 2700 versch. Positions-Bezeichnungen,
- Darunter 143 Schreibweisen für „Geschäftsführer“
 - Abkürzungen: GF, GSF, GeschF
 - Tippfehler: Geschfätsführer
 - Genauigkeit: Geschäftsführerin
- Gleiches Problem bei Branche
 - Maschinenbau
 - Bau von Werkzeugmaschinen

Unstrukturierte Inhalte

- Prävention: Vorgaben, Auswahlen
- Lösung für Altdaten: Vereinheitlichen
 - oder Anreichern aus externen Quellen

Inhalte in den falschen Feldern

- Hauptsächlich ein Problem der Adresse selbst
- Für Analyse weniger relevant, aber für Aktionen
 - „Sehr geehrter Herr Einkaufsabt.“

Veraltete Daten

- Unterschiedlich schwerwiegend,
- viele Analyse-Prädiktoren veralten eher langsam
 - B2B: Branchencode, Gründungsjahr
 - B2C: Sexcode, Geburtsjahr, Familienstatus
- Prävention: keine
- Lösung: immer wieder abfragen, extern anreichern

Fehlende Daten

- Oft ein Problem der Motivation, die Information zu erfragen
- Oder auch nur die bekannte Information einzutragen

Falsche Inhalte

- Fehl-Erfassung
 - Z.B. 3% Sexcode-Fehler wg. Akkord-Bezahlung
- Absicht
 - Z.B. im Web
- Prävention:
 - Motivation für richtige Erfassung geben
 - Technische Hilfestellung per Redundanz-Check und Rückmeldung

Lösung für Vergangenheit

- Anhand von Redundanz fehlerhafte Sätze aufspüren
- die vermutlich fehlerhaften Sätze bei der Analyse ausmaskieren oder sonderbehandeln
- Manche Daten können gleich korrigiert werden
 - Z.B. Sexcode anhand von Vorname

Dubletten

- ***Was sind Dubletten?***
- Für die Analyse = zusammengehörige Daten verteilt auf unzusammenhängende Datensätze

Abweichungen in der Schreibweise/Privatadressen

- Tippfehler: Alexander wird zu Alexnader
- Schreibfehler: Chrychanowski wird zu Krischanoffsky
- verschiedene Namens-Schreibweisen: Meier/Mayer, Sylvia/Silvia, Mathias/Mattias
- Freud'sche Hörfehler: Sponheimstr. wird zu Sponhainstr.
- Freud'sche Fehler: Friedrich-Silcher-Weg wird zu Friedrich-Schiller-Weg, weil das Gehirn schon bei "Friedrich" auf "Schiller" schaltet.

Abweichungen in der Schreibweise/Firmenadressen

- Vertauschungen: Müller Möbel / Möbel-Müller
- Zusätze: Müller Möbel / Müller Möbel-Fabrikation GmbH
- Wort-Mutationen: Ranitzki Soft- und Hardware / Ranizki Hard + Software
- Abkürzungen: Westf. Inst. f. Agrarökonomie / Agrarökonomisches Institut Westfalen
- Feld-Vertauschungen



Bessere Daten - mehr Erfolg

Die paar Dubletten?

- Ohne hochwertige Massnahmen:
 - 4-6% Dubletten selbst in gut gepflegten Datenbanken
 - 30% (!) Dubletten bei den Stammkunden
 - da Stammkunden die wichtigste Analysegruppe sind, werden Auswertungen signifikant verfälscht.
- Im B2B-Bereich trotz Einsatz von „bisher ausreichender“ Software oft 1-2% Restdubletten

Klassische Folgen von Dubletten

- Doppelte Mailing-Kosten
- Schlechter Eindruck: Massenwerbung

Folgen speziell im aCRM

- Analyse-Verfälschungen
 - Warenkorb-Analysen
 - Anstossketten-Analysen
 - CLV
- Aktions-Fehler
 - Z.T. peinliche
 - Stammkunde „Küchenbau Ranitzky“ erhält 15% Treuerabatt
 - Vermeintlicher Neukunde „Ranizki Küchen- und Innenausbau“ erhält 25% Frühjahrsrabatt

Was kann man tun?

- Entscheidend: Prävention!
- Wenn Dubletten gleich bei der Eingabe erkannt werden,
 - kann der Mitarbeiter nachfragen oder nachlesen (Höhere Klärungsrate)

Nachträgliche Zusammenführung ist teuer, denn

- Unklare Dubletten bleiben oder verursachen hohen Klärungsaufwand
- Widersprüchliche Informationen: Oft weiss man nicht, was richtiger ist
- Die Zusammenführung der Infos muss teilweise manuell erfolgen

**Vielen Dank für Ihre
Aufmerksamkeit!**