

# Big Data Anwendungen

---

Datenqualität

# Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
  - Fehlende Werte
  - Unreine Daten
  - Transformation von Daten
  - Externe Datenquellen
  - Sampling
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

# Fehlende Werte

- Beschreibung:
  - Einzelne Attribute nicht angegeben (technisch: „null“-Werte)
  - Attributwerte sind offensichtlich falsch (Alter > 150 Jahre)
- Ursachen (Auswahl)
  - Nicht vorhandene oder unvollständige Angaben durch Befragten (z.B. in Fragebögen, freiwillige Angaben in Webformularen)
  - Softwareupdate liefert neue Daten (z.B. Monitoring weiterer Sensoren, Tracking von Sendungsnummern)
  - Abhängige Daten wurden gelöscht (z.B. Löschung aus Datenschutz, Artikelnummern wiederholt vergeben)
- Probleme
  - Analysealgorithmen können nicht mit fehlenden Werten umgehen
  - Interpretation erfordert oft gutes Verständnis der Daten
  - Unterscheidung zwischen „Weiß-nicht“ und fehlenden Werten oft schwer
- Anmerkungen
  - Behandlung von Fehlwerten zur „Rettung“ von Beobachtungen
  - Behandlung von Fehlwerten bringt Risiko systematischer Fehler

# Eliminierungsstrategien

- Eliminierung von Beobachtungen
  - Entfernung von Beobachtungen mit mindestens  $n$  fehlenden Attributen
  - Vorteile
    - Manchmal einzige Lösung: z.B. wenn Attribut zentral
    - Strategie ist einfach
  - Nachteile
    - Erhebliche Reduktion der Datenmenge  
(zusätzliche Gefahr: nur ohnehin bekannte Kunden übrig)
    - Oft haben fehlende Wert inhaltliche Bedeutung  
(z.B. Beruf: „Hausfrau“, GfK-Stile, Temperaturbereich von Sensoren)
- Eliminierung von Attributen
  - Verzicht auf unvollständig beobachtete Merkmale
  - Vorteil: Strategie ist einfach
  - Nachteile
    - Ausschluss wichtiger Informationen möglich
    - Strategie motiviert zu oberflächlicher Herangehensweise

# Imputation

- Ersetzen fehlender Werte durch „vernünftige“ Werte
- Prinzipiell verschiedene Verfahren
  - Einfache Imputation
    - Fehlende Werte durch einen Wert ersetzt
    - Problem: Unterschätzung der Varianz in den Daten
  - Multiple Imputation
    - Ersetzen durch mehrere Imputationswerte
    - Kombination der gewichteten Imputationswerte
    - Nachteile
      - Zeitaufwendig
      - Sensitiv bzgl. Gewichtung
  - Imputation mit Vorhersagemodell
    - Modell zur Vorhersage des Wertes...  
... auf Basis anderer Werte der Beobachtung
    - Nachteil
      - Zeitaufwendig
      - Eigenes Modell muss erstellt werden

# Einfache Imputation – Manuelle Verfahren

- Manuelles Auffüllen
  - Ein Experte prüft Beobachtungen und pflegt Daten nach
  - Nachteile
    - Hoher manueller Aufwand  
(Bsp. Nachpflegen von GfK-Stilen)
    - Auffüllender muss über Expertenwissen verfügen
- Ersetzen durch globale Konstante
  - Alle fehlenden Werte durch „unbekannt“ oder „-∞“ ersetzt
  - Nachteil: Funktioniert nur, wenn fehlender Wert eindeutige Ursache
  - Gegenbeispiele
    - Unternehmen ersetzt fehlendes Alter durch 17
    - Temperaturfühler fällt bei zu hohen und zu geringen Temperaturen aus
- Abbilden der Information über „neues“ Attribut
  - Inhalt eines Attributs wird über andere Attribute ermittelt
  - Beispiele
    - Modizität eines Artikel über Alter des durchschnittlichen Kunden
    - Saisonalität eines Artikels über Nachfrageverlauf im Jahr

# Einfache Imputation – Standardverfahren

- Einsetzen des Mittelwertes
  - Mittelwert wird über nicht fehlende Werte ermittelt
  - Nachteile
    - Nur bei metrischen Attributen sinnvoll
    - Bei zu vielen Fehlwerten: Verteilung wird sehr „spitz“
- Einsetzen des Modalwerts
  - Modalwert wird über nicht fehlende Werte ermittelt
  - Nachteile:
    - Schwierig bei bi- oder multimodalen Attributen
    - Kann zu Überrepräsentation eines Werts führen
- Einsetzen des Modal- oder Mittelwertes der Klasse
  - Kunden werden einer Klasse zugeordnet
  - Mittelwert wird über nicht fehlende Werte der Klasse ermittelt
  - Nachteile
    - Wahl der „Klassifikation“ muss mit Vorhersage korrelieren

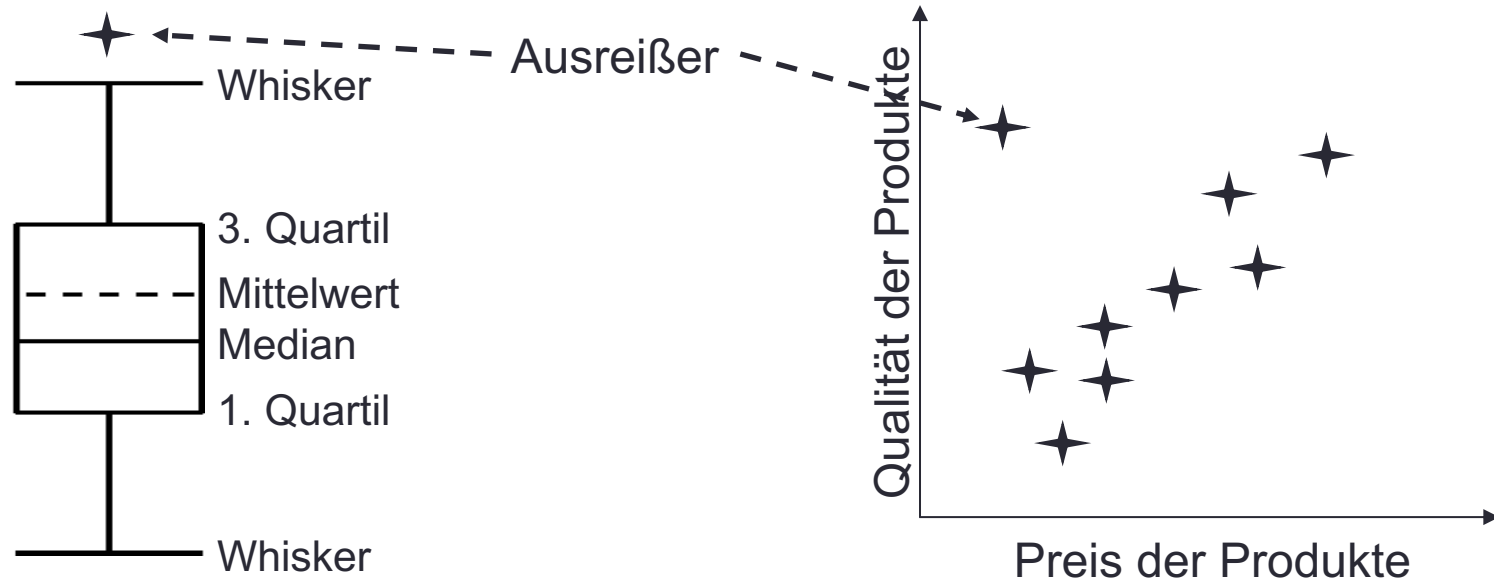
# Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
  - Fehlende Werte
  - Unreine Daten
  - Transformation von Daten
  - Externe Datenquellen
  - Sampling
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte



# Unreine Daten – Ausreißer

- Beschreibung
  - Ausreißer unterscheiden sich in der Lage stark von anderen Datenpunkten
  - Oft sind Ausreißer inhaltlich „korrekt“, ...  
... haben aber negativen Einfluss auf statistische Analyse
- Idee: Vermeidung von Ausreißern
  - Auswertung nur auf 90% der zentralen Daten
  - Boxplots mit „Whiskers“ (Ausschluss der oberen/unteren 5%)



# Unreine Daten – Beispiele

Fehlerart	Beispiel	Anmerkungen
Ausreißer	monats_gehalt=1.00 €	Prüfung nötig (ggfs. korrekte Werte)
Tippfehler	stadt="Maagdeburg"	phonetische Fehler, Verständnisfehler
Abkürzungen	erfahrung="A"	
Eingebettete Werte	name="A. Maier 12.01.1962"	mehrere Werte in einem Attribut
Fehleinordnungen	stadt="Deutschland"	
Inkonsistente Werte	stadt="Magdeburg" plz=53212	Stadt und PLZ passen nicht
Vertauschungen	name <sub>1</sub> ="M. Müller" name <sub>2</sub> ="Schmidt P."	Typischerweise in Freiformfeldern
Duplikate	name <sub>1</sub> ="Max Müller" name <sub>2</sub> ="Max Mueller" name <sub>3</sub> ="M. Müller"	Inkonsistente Eingaben Fehler in der Eingabe
Widersprüche	alter="90" geb.datum="01.01.1999"	
Ungültige Verweise	abteilung=20	Abteilung existiert nicht

# Unreine Daten – Kölner Phonetik

- Idee
  - Ähnlich klingende Buchstaben...  
... mit identischer Zahl kodiert
  - Entfernung von...
    - ... „Nicht“ klingenden Buchstaben
    - ... Doppelbuchstaben
    - ... Vokalen
- Beispiele
  - Relevante Codes  
M  $\Rightarrow$  6, U/Ü  $\Rightarrow$  0, L  $\Rightarrow$  5, E  $\Rightarrow$  0, R  $\Rightarrow$  7
  - Kodierung von Müller
    - Kodierung: Müller  $\Rightarrow$  605507
    - Entfernung Dup: 605507  $\Rightarrow$  60507
    - Entfernung 0: 657
  - Kodierung von Mueler
    - Kodierung: Mueler  $\Rightarrow$  600507
    - Entfernung Dup: 600507  $\Rightarrow$  60507
    - Entfernung 0: 657

Buchstabe	Kontext	Code
A, E, I, J, O, U, Y		0
H		–
B		1
P	nicht vor H	
D, T	nicht vor C, S, Z	2
F, V, W		3
P	vor H	
G, K, Q		
C	<u>im Anlaut vor A, H, K, L, O, Q, R, U, X</u> vor A, H, K, O, Q, U, X außer nach S, Z	4
X	nicht nach C, K, Q	48
L		5
M, N		6
R		7
S, Z		
C	<u>nach S, Z</u> <u>im Anlaut außer vor A, H, K, L, O, Q, R, U, X</u> nicht vor A, H, K, O, Q, U, X	8
D, T	vor C, S, Z	
X	nach C, K, Q	

# Unreine Daten – Berücksichtigung bei Erhebung

- Ursache vieler Verunreinigungen: Menschliche Fehler
- Alternative 1: Technische Maßnahmen zur Fehlerreduktion
  - Eingabemasken erlauben nur „gültige“ Werte  
(z.B. Prüfung von E-Mail-Adressen auf Format [Text]@[Text].[Text])
  - Vergleich verschiedener Felder  
(z.B. Abgleich PLZ, Straße und Stadt – Anekdote: Wo wohnt Stephan?)
  - Angebot von Auswahlfeldern  
(z.B. Wahl des Bundeslands)
- Alternative 2: Etablierung von Prüfprozessen
  - Prüfung der Datenqualität durch verschiedene Entscheider  
(z.B. Preispflege und Preisfreigabe durch separate Mitarbeiter)
  - Algorithmische Plausibilitätsprüfung der Eingaben...  
... und Prüfung von unplausiblen Werten durch Experten  
(z.B. Nachfrage an Domänenexperten bei Ausreißern)

# Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
  - Fehlende Werte
  - Unreine Daten
  - Transformation von Daten
  - Externe Datenquellen
  - Sampling
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

# Skalierung

- Wünschenswert  
Alle Attribute haben für Vergleichbarkeit ähnlichen Wertebereich

- Standardisierung

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Normierung mit Mittelwert und Standardabweichung
- Vorteil
  - Richtung der Abweichung vom Mittelwert ablesbar
  - Wenig sensitiv bezüglich Ausreißern
- Min-Max-Skalierung

$$z_i = \frac{x_i - \min_{\forall j} x_j}{\max_{\forall j} x_j - \min_{\forall j} x_j}$$

- Normierung mit Minimum und Maximum der Daten
- Vorteil
  - Wertebereich sicher [0...1]
  - Werte leicht verständlich

# Binning

- Problem
  - Daten oft sehr feingranular
  - Wesentliche Information grobgranular
- Idee
  - Kombinieren der Werte in Gruppe
  - Zuordnung von...  
... repräsentativem Wert
- Ansätze
  - Ersetzen durch Mittelwert
  - Ersetzen durch Median
  - Ersetzen durch Binmitte
- Vorteile
  - Geringeres Datenvolumen
  - Variable Bingröße erlaubt Abbildung multimodaler Werte

Uhrzeit	Temperatur	Gebinnt	Wert
06:00	18.01°	]15,20]	17.5°
07:00	20.45°	]20,25]	22.5°
08:00	22.67°	]20,25]	22.5°
09:00	24.76°	]20,25]	22.5°
10:00	26.23°	]25,30]	27.5°
11:00	27.05°	]25,30]	27.5°
12:00	28.00°	]25,30]	27.5°

# Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
  - Fehlende Werte
  - Unreine Daten
  - Transformation von Daten
  - Externe Datenquellen
  - Sampling
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte



# Das Problem

- Traditionell (bis zur Jahrtausendwende)
  - „Kundenbestand“ ist Kapital des Versandhändlers
  - Über Scoring Identifikation der „besten Kunden“
    - Kunden mit geringer Retourenquote
    - Kunden mit hoher Bestellfrequenz
    - Kunden mit hohem Bestellwert
  - Versand von Werbemitteln an „beste Kunden“
- Aktuell
  - Fluktuation der Kunden zwischen Händlern hoch
    - Hohe Preisaffinität
    - Geringe Händlerbindung / starke Markenbindung
  - Bedeutung des „Kundenbestands“ sinkt
  - Zielgerichtete Ansprache immer „neuer“ Kunden

# Die Lösung

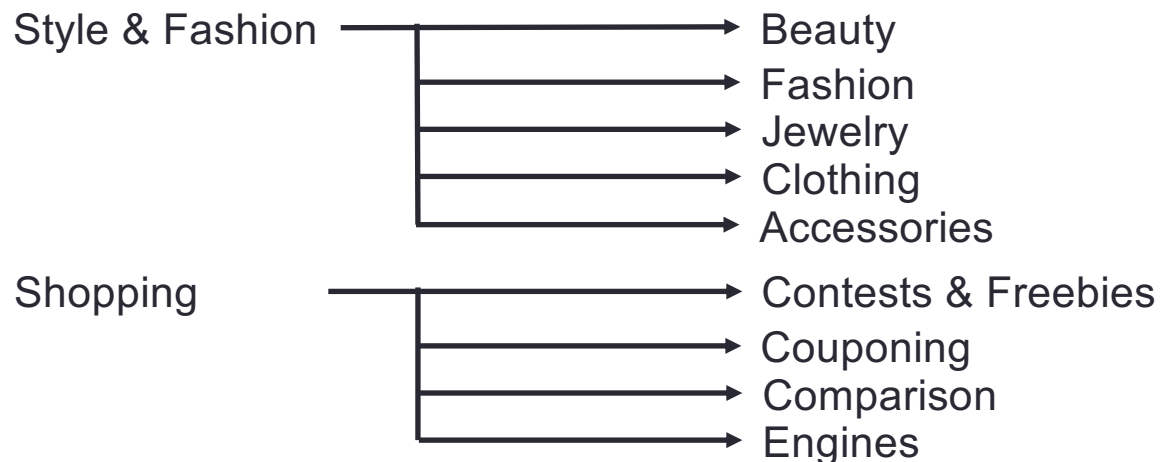
- Data Management Plattformen (DMPs)
  - Kaufen Daten aus unterschiedlichen Quellen
    - Postalische / E-Mail-Adressen
    - Soziodemographische Daten
    - Kaufkraft
    - Affinität für Produktgruppen
    - Shopping-Portal-Daten
  - Bringen Daten aus den Quellen zusammen
  - Bauen mit Daten „Taxonomien“ auf
  - Können Daten des Händlers mit eigenen Daten matchen
- Mit zunehmender Fluktuation der Kunden...
  - ... steigt die Bedeutung der DMP...
  - ... gegenüber dem Kundenbestand

# Ontologie

- Teildisziplin der Philosophie
  - „Die Lehre von dem, was ist.“
- In der Informatik (Gruber, 1995)
  - „Eine Ontologie ist eine explizite Spezifikation einer Konzeptualisierung eines Anwendungsbereiches.“
- Problem:
  - Oft unüberschaubare Anzahl von Fachbegriffen und Synonymen
  - Kommunikationsschwierigkeiten unter Menschen / Systemen
- Idee: Standardisierung
  - ... der Begriffe unter den Benutzern
  - ... des Verständnisses der Begriffe
- Fokus hier: Taxonomien („Ontologie mit Baumstruktur“)

# IAB Tech Lab Content Taxonomy

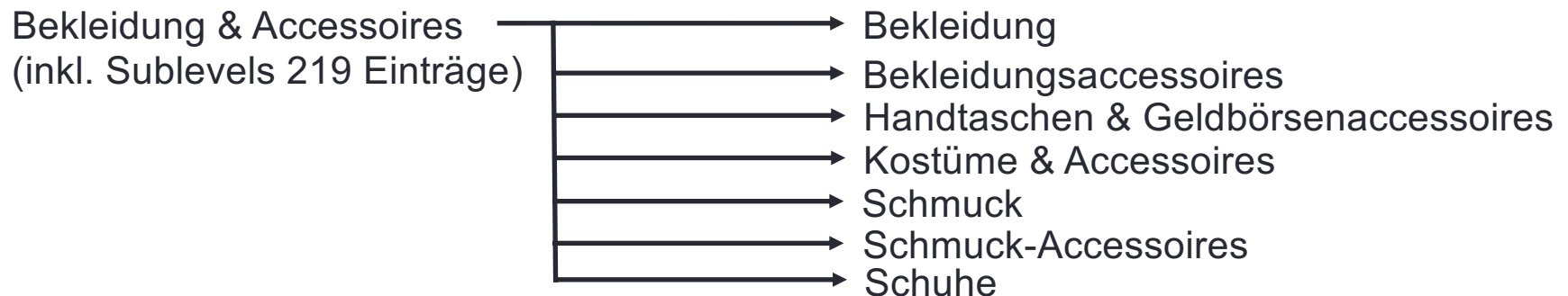
- Aufbau der Taxonomie
  - Zwei Ebenen: Tier 1, Tier 2
  - Anzahl Einträge: 361 (auf Tier 2)
- Für Kleidungshändler relevant



- Weitere Bereiche  
Religion, Reisen, Sport, Haustiere, Haus & Garten, ...

# Google Product Taxonomy

- Aufbau der Taxonomie
  - Fünf Ebenen: Abhängig vom Oberbegriff meist nur 2 besetzt
  - Anzahl Einträge: 5427 (davon ca. 200 auf Ebene 2)
- Für Kleidungshändler relevant



- Weitere Bereiche  
Baby & Kleinkind, Fahrzeuge & Teile, Für Erwachsene,  
Gesundheit & Schönheit, Haus & Garten, Möbel,  
Nahrungsmittel, Getränke & Tabak, Sportartikel, ...

# Zentrale Aufgabe bei der Datenanreicherung

- Beobachtung der Taxonomie für jeden Kunden und jeden Tag...  
... Änderungen müssen erkannt werden
- Beispiele
  - Kunde wechselte letzte Nacht bei Bekleidung von 0 auf 1:
    - Entweder hat Kunde nach Bekleidung gesucht...  
... oder sogar gekauft
    - Wechsel von 0 auf 1 und Zeitpunkt hat also deutlich...  
... mehr Aussagekraft als „bloßer“ Zustand
  - Kunde wechselte letzte Nacht bei Couponing von 1 auf 0:
    - Untersuchung auf Gesetzmäßigkeiten möglich...  
... damit besseres Verständnis des DMP Algorithmus
- Ausnahme  
Soziodemographische Daten (Geschlecht, ...)

# Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
  - Fehlende Werte
  - Unreine Daten
  - Transformation von Daten
  - Externe Datenquellen
  - Sampling
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

# Stichprobe

- Idee
  - Algorithmen zur Datenanalyse oft ressourcenintensiv...  
... mit dem Ziel „charakteristische“ Eigenschaften zu identifizieren
  - Ziehen einer repräsentativen Stichprobe
- Ansätze
  - Biased Sampling
    - Jüngere Beobachtungen werden eher für Stichprobe gezogen
    - Vorteile
      - Sample repräsentiert aktuellen Kunden, ... gut
      - Historische Daten werden nach und nach verdrängt
  - Stratified Sampling
    - Separierung der Daten in Klassen und ziehen für Klassen einzeln
    - Vorteile
      - Analysen auch auf Klassen mit wenigen Beobachtungen möglich (z.B. SPAM Klassifikation von E-Mails)
      - Einfach zu realisieren