Normalformen

Zweiter Schritt der Zerlegung der Relation *Filmliste* basierend auf der FD :

Kino → Telefonnummer (Tel.)

Kino-Screens

Kino	Tel.	Zeit	Titel
Cine. City	441111	11:30	The Hobbit
Cine. City	441111	14:30	The Lord of the Rings 3
Odeon	442222	11:30	Adventures of Tintin
Odeon	442222	14:00	War Horse
Odeon	442222	16:30	The Lord of the Rings 3

Filme

Titel	Regisseur
The Hobbit	Jackson
The Lord of the Rings 3	Jackson
Adventures of Tintin	Spielberg
War Horse	Spielberg

Kino

Kino	Tel.
Cine. City	441111
Odeon	442222

Screens

Kino	Zeit	Titel
Cine. City	11:30	The Hobbit
Cine. City	14:30	Saving Private Ryan
Odeon	11:30	Adventures of Tintin
Odeon	14:00	War Horse
Odeon	16:30	Saving Private Ryan

Abhängigkeitsbewahrung

- Idee: Alle FDs, die für die Relation R gelten, sollen *lokal* auf den Relationen $R_1, \ldots, R_n$ aus der Zerlegung überprüfbar sein.
- Die Projektion von F auf α (F_α) ist die Menge von FDs aus F^+ die nur Attribute aus α enthalten:

$$F_\alpha = \{ \beta \to \gamma \in F^+ \mid \beta\gamma \subseteq \alpha \}$$

Abhängigkeitsbewahrung

- Berechnung von FD Projektionen :

```
Input: F, α
Output: F_α
Erg = ∅
For each β ⊆ α do
    T = β⁺ (bzgl. F)
    Erg = Erg ∪ {β → T ∩ α}
Return Erg
```

Hüllentreue Zerlegung

- **Definition.**

  Die Zerlegung der Relation R in Relationen $R_1,\dots, R_n$ wird als hüllentreue Zerlegung bezeichnet falls:

$$F_R^+ = (F_{R_1} \cup \ \dots \cup F_{R_n})^+$$

- Intuitiv, müssen $F_{R_1} \cup \ \dots \cup F_{R_n}$ und F_R äquivalent sein

Zerlegungbeispiel

Fun fact: In Deutschland gibt es über 2000 Goethestraßen!

- PlzVerzeichnis(Straße, Ort, BLand, PLZ)
- Bedingungen:
  - Orte werden durch *Ort* und *BLand* eindeutig charakterisiert
  - Innerhalb einer Straße ändert sich *PLZ* nicht
  - PLZ-Gebiete gehen nicht über Ortsgrenzen, Orte nicht über Bundeslandgrenzen
- FDs: {PLZ} → {Ort, BLand} und {Straße, Ort, BLand} → {PLZ}
- Die Zerlegung {PLZ, Straße} und {PLZ, Ort, BLand} ist:
  - Verlustlos (Corollary)
  - Nicht abhängigkeitserhaltend (die zweite FD ist *lokal* nicht überprüfbar)

Zerlegung einer Relation

Ist die Relation mit Blick auf FDs redundanzfrei?

Keine Zerlegung notwendig

Starte Zerlegungsprozedur

Ist die Zerlegung der Relation R in R_1 und R_2 verlustlos möglich?

Keine Zerlegung möglich

Ist die Aufspaltung der zugehörigen FDs in FD_1s und FD_2s, die jeweils R_1 und R_2 zugeordnet werden könnten, abhängigkeitserhaltend möglich?

Ist ok, aber „unschön", da Überprüfung der FDs nur nach der Rekonstruktion von R möglich ist (Effizientverlust)

Normalisierung

- Normalformen definieren Qualitätskriterien (Vermeiden von Inkonsistenzen)
- Redundanz ist oft die Ursache von Schemaproblemen (keine FDs, keine Redundanz)
- Normalisierung:
  - Jede Relation entspricht genau einer Objektmenge oder genau einer Relationship-Menge zwischen Objekten
  - Alle Informationen und Integritätsbedingungen (Schlüsselkandidaten und Fremdschlüssel) sind abgebildet
  - Redundanz ist weitgehend eliminiert
  - Es treten keine Änderungsanomalien auf

  ⇒ **Beseitigung von Abhängigkeiten innerhalb der Relation**

Normalisierung

- Integritätsregeln, insbesondere FDs, helfen ein schlechtes Schema zu identifizieren
- Die Lösung ist meistens die Zerlegung der Relation
- **Fragestellung**: wann ist eine Zerlegung möglich, und wann ist sie notwendig?

Normalformen

- Wenn eine Relation in einer Normalform ist, dann wissen wir, dass bestimmte Probleme nicht vorkommen können.
- Es hilft, um zu bestimmen ob eine Zerlegung weiter hilft oder nicht.
- Die Normalformen basierend auf FDs: 1NF, 2NF, 3NF, BCNF

$$\textbf{BCNF} \subseteq \textbf{3NF} \subseteq \textbf{2NF} \subseteq \textbf{1NF}$$

1. Normalform

- **Definition.** Eine Relation ist in der ersten Normalform, wenn alle Attribute der Relation **atomar** sind. Zusammengesetzte und mehrwertige Attribute sind nicht erlaubt.
- Nicht atomar: Dictionaries, Listen, Untertabellen, json- o.ä. Strukturen, usw
- Ist bei der benutzten Definition des relationalen Modells normalerweise automatisch eingehalten
- Atomar: String, Integer, etc.
- Bemerkung: Datum / Uhrzeit wird als atomar betrachtet
- Was ist atomar genug?
    - (Name, Vorname) oder (Name und Vorname)
    - (Jahr, Monat, Tag, Stunde, Minute, Sekunde) oder (Timestamp)

Relation nicht in 1NF

Kunden (<u>*KundenID*</u>, *Name, Vorname, Produkte*)

Transformation in 1NF durch Zerlegung

Kunden (<u>*KundenID*</u>, *Name, Vorname*)
Bestellung *(* <u>*KundenID, Produkt*</u>)

Referenzielle Integritätsbedingung (Fremdschlüssel):
Bestellung.*KundenID* referenziert Kunden.*KundenID*

<u>KundenID</u>	**Name**	**Vorname**	**Produkte**
K1	Nuhr	Dieter	Asus
K2	Pelzig	Erwin	iPhone, iPad, Vaio
K3	Gruber	Monika	Nexus, Galaxy Tab, Kindle

<u>KundenId</u>	**Name**	**Vorname**
K1	Nuhr	Dieter
K2	Pelzig	Erwin
K3	Gruber	Monika

<u>KundenID</u>	**<u>Produkt</u>**
K1	Asus
K2	iPhone
K2	iPad
K2	Vaio
K3	Nexus
K3	Galaxy Tab
K3	Kindle

2. Normalform

- **Definition.** Eine Relation R mit zugehörigen FDs F ist in zweiter Normalform, genau dann wenn es in 1NF ist und **jedes Nichtschlüssel-Attribut** $A \in R$ voll funktional abhängig ist von **jedem Kandidatenschlüssel** der Relation.

- Eine funktionalle Abhängigkeit $X \rightarrow Y$ heißt **voll**, wenn es keine echte Teilmengen $Z \subset X$ gibt, s.d. gilt $Z \rightarrow Y$

- Intuitiv: Ein Relationenschema verletzt 2NF, wenn in der Relation Informationen über mehr als ein einziges Konzept modelliert werden.
- Selbst bei Erfüllung der 2NF können immer noch Redundanzen im Schema enthlaten sein (durch transitive Abhängigkeiten)

2NF - Beispiel

- StudentenBelegung (MatrNr, VorlNr, Name, Semester)
- Primärschlüssel: {MatrNr, VorlNr}
- FDs welche 2NF verletzen:
  - {MatrNr} → {Name}
  - {MatrNr} → {Semester}
- **Bemerkung**. Hier können wir auch die Anomalien wiedererkennen.
- Zerlegung:

  Enrolled(MatrNr, VorlNr) - vllt. auch Note

  Studenten(MatrNr, Name, Semester) - vllt. auch Email, Geburtsdatum, u.a.

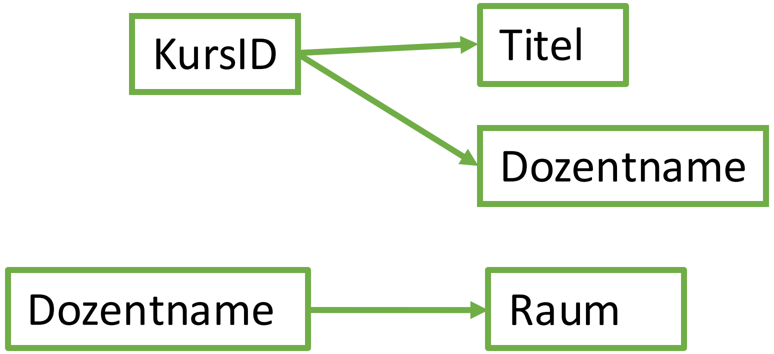
3. Normalform

- **Definition.** Eine Relation mit zugehörigen FDs F ist in der 3. Normalform, wenn für alle Abhängigkeiten $A \rightarrow B$ aus F^+ gilt:
  - $B \subseteq A$ (FD ist trivial) **oder**
  - A enthält einen Schlüssel von R (A ist ein Superschlüssel) **oder**
  - B ist Teil eines Schlüsselkandidaten (B ist prim)
- **Äquivalente Def.** Eine Relation R ist in 3NF, wenn sie in der 2NF ist und **kein Nichtschlüsselattribut von einem Schlüsselkandidaten** transitiv abhängt.
- **Bemerkung.** 3NF beseitigt Abhängigkeiten von Nicht-Schlüsselattributen.
- Selbst bei Erfüllung der 3NF sind Redundanzen möglich.
- Eine Relation bleibt in 3NF, wenn BCNF nicht erreichbar ist (keine gute Zerlegung oder aus Leistungsgründen)

Relation nicht in 3NF

- Kurs (KursID, Titel, DozentName, Raum) - nicht in 3NF

- Es gilt: {DozentName} → {Raum}
- **Aber**: Dozentname ist kein Schlüssel und Raum ist nicht Teil eines Schlüsselkandidaten
- Folgende Anomalien können auftreten:
  - Dozenten und Raum sind ohne Zuordnung eines Kurses nicht vefügbar
  - Falls ein Dozent keinen Kurs liest, werden alle Informationen über den Dozenten und seinem Raum gelöscht
- Schema in 3NF:
  Kurs (KursID, Titel, DozentName)
  Dozent (DozentName, Raum)



Boyce-Codd Normalform

- **Definition.** Eine Relation R mit zugehörigen FDs F ist in der Boyce-Codd Normalform, wenn für alle Abhängigkeiten $A \rightarrow B$ aus F^+ gilt:
  - $B \subseteq A$ (FD ist trivial) **oder**
  - A enthält einen Schlüssel von R (A ist ein Superschlüssel)
- Die BCNF hat also außer den trivialen nur noch funktionalen Abhängigkeiten deren Determinante (linke Seite) ein Superschlüssel ist
- **Bemerkung.** Wenn R in BCNF ist, ist es automatisch auch in 3NF

Relation in 3NF aber nicht BCNF

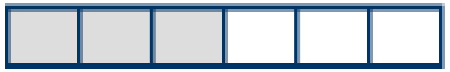
- Städte (*Ort, BLand, Ministerpräsident, Einw*)
- Schlüsselkandidaten:
  - (*Ort, BLand*)
  - (*Ort, Ministerpräsident*)
- FDs:
  - *BLand* → *Ministerpräsident*
  - {*Ort, BLand*} → *Einw*
  - *Ministerpräsident* → *BLand*
- Relation ist in **3NF**, aber **nicht** in **BCNF**
- Anomalien können auftreten, da die Information, wer welches Bundesland regiert, mehrfach abgespeichert wird

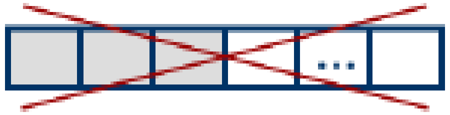
Normalformen - Übersicht

- 1. Normalform (1NF)
  - Keine mehrwertigen Attribute
- 2. Normalform (2NF)
  - Keine Vermischung von Sachverhalten in Relationen
- 3. Normalform (3NF)
  - Keine funktionalen Abhängigkeiten von Nichtschlüsselattributen innerhalb von Relationen
- Boyce-Codd-Normalform (BCNF)
- Weitere Normalformen
  - 4. Normalform (4NF)
  - 5. Normalform (5NF)

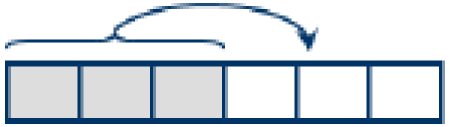
Normalformen - Zusammenfassung

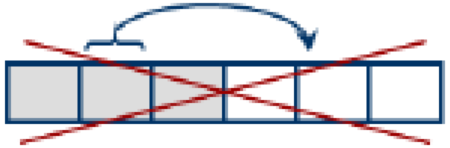
1NF – alle Attribute sind atomar



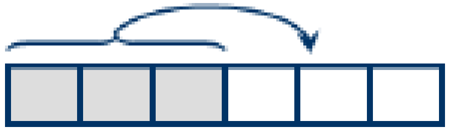


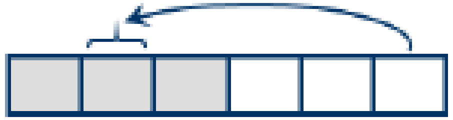
2NF – alle Nichtschlüsselattribute sind voll funktional abhängig von jedem Kandidatenschlüssel (keine partiellen Abhängigkeiten)

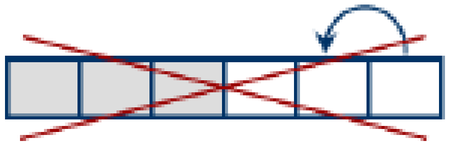




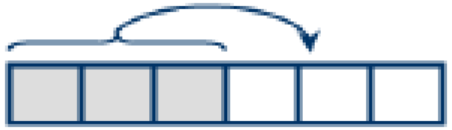
3NF – in 2NF und alle Nichtschlüsselattribute sind nur von Kandidatenschlüssel abhängig (keine transitiven Abhängigkeiten)

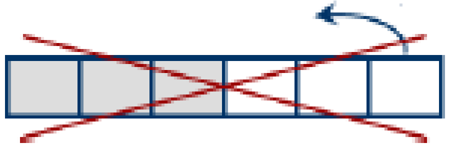






BCNF – jede Determinante ist ein Superschlüssel (alle FDs werden von Kandidatenschlüsseln bestimmt)





3NF vs. BCNF

- **Bemerkung.** Man kann jede Relation R so in $R_1,\ldots, R_n$ zerlegen, dass gilt:
  - Die Zerlegung ist **verlustlos** und **abhängigkeitsbewahrend**
  - R_i ist in **3NF**, $1 \leq i \leq n$
- **Bemerkung.** Man kann jede Relation R so in $R_1,\ldots, R_n$ zerlegen, dass gilt:
  - Die Zerlegung ist **verlustlos**
  - R_i ist in **BCNF**, $1 \leq i \leq n$
- Aber man kann nicht immer eine BCNF-Zerlegung finden, die auch abhängigkeitsbewahrend ist.

BCNF-Zerlegung

- Wenn die funktionale Abhängigkeit $\alpha \rightarrow \beta$ die BCNF verletzt, dann können wir die Relation in $R - \beta$ und $\alpha \cup \beta$ zerlegen. Wir können das weitermachen bis alle neuen Relation in BCNF sind (es geht immer zu Ende) (Zerlegung Korollar).
- Diese Zerlegung wird verlustlos, aber nicht unbedingt abhängigkeitsbewahrend sein
- **Bem**. Wenn es mehrere Abhängigkeiten gibt welche die BCNF verletzten, dann macht es einen Unterschied welche wir als erste für die Zerlegung auswählen.

BCNF – Zerlegungsbeispiel

- R(<u>C</u>, S , J, D, P, Q, V)
- F ={ JP $\rightarrow$ C, SD $\rightarrow$ P, J $\rightarrow$ S}

- SD $\rightarrow$ P verletzt BCNF $\Rightarrow$ Zerlegung (<u>S,D</u>,P) mit F_1 = {SD $\rightarrow$ P} und
  (<u>C</u>,S,J,D,Q,V) mit F_2 = {J $\rightarrow$ S}

- J $\rightarrow$ S verletzt BCNF für die zweite Relation $\Rightarrow$ (<u>C</u>,S,J,D,Q,V) wird in
  (<u>J</u>, S) und (<u>C</u>, J, D, Q, V) zerlegt
- D.h. R wird in (<u>S,D</u>,P) , (<u>J</u>, S) und (<u>C</u>, J, D, Q, V) zerlegt – verlustlos, aber nicht abhängigkeitsbewahrend

Kanonische Überdeckung

- **Definition (Äquivalenz funktionaler Abhängigkeiten).**

  Zwei Mengen F und G von FDs eines Relationenschemas R sind äquivalent, falls $F^+ = G^+$ gilt.
- Wunsch: Berechne eine möglichst kleine Menge, die zu F äquivalent ist → geringer Aufwand beim Testen, ob ein neues Tupel eine FD verletzt

Kanonische Überdeckung

- Man kann **überflüssige Attribute** durch Links- und Rechtsreduktion entfernen:
- **Linksreduktion**
  - Führe für jede FD $A \rightarrow B$ aus F die Linksreduktion durch, indem für alle $X \in A$ überprüft wird ob das Attribut X überflüssig ist, d.h. ob gilt

    $B \subset$ Hülle(F, A - {X}) , also $(A - \{X\})^+$ in Beziehung zu F

    (anders gesagt $(F - \{A \rightarrow B\} \cup \{(A-\{X\}) \rightarrow B\})^+ = F^+$)

    Ist dies der Fall, ersetze $A \rightarrow B$ durch $A - \{X\} \rightarrow B$
- **Rechtsreduktion**
  - Führe für jede FD $A \rightarrow B$ aus F die Rechtsreduktion durch, indem für alle $Y \in B$ überprüft wird ob das Attribut Y überflüssig ist, d.h. ob gilt

    $Y \in$ Hülle(F – (A → B) ∪ (A → B – {Y}), A)

    (anders gesagt ($F - \{A \rightarrow B\} \cup \{A \rightarrow (B - \{Y\})\})^+ = F^+$)

  Ist dies der Fall, ersetze $A \rightarrow B$ durch $A \rightarrow B - \{Y\}$

Kanonische Überdeckung

- **Definition (Kanonische Überdeckung).**

  Die Menge von FDs F_c wird als **kanonische Überdeckung** von F bezeichnet, falls folgende Bedingungen erfüllt sind:
  - $F_c^+ = F^+$ (Äquivalenz) (1)
  - Für alle FDs A → B in F_c gibt es keine überflüssigen Attribute in A und in B, d.h.
    - für alle Attribute C aus A gilt (F_c - {A → B} ∪ {(A-{C}) → B})⁺ $\neq F^+$ (2)
    - für alle Attribute D aus B gilt (F_c - {A → B} ∪ {A → (B – {D})})⁺ $\neq F^+$ (3)
    - jede linke Seite der FDs in F_c kommt nur einmal vor, d.h. Falls

      A → B und A → C, dann wird in F_c nur die FD A → B ∪ C vrewendet (4)
- Intuitiv: (2) – keine überflüssigen Attribute auf der linken Seite von FDs

  (3) - keine überflüssigen Attribute auf der rechten Seite von FDs