TU München, Fakultät für Informatik
Lehrstuhl III: Datenbanksysteme
Prof. Alfons Kemper, Ph.D.

Übung zur Vorlesung *Grundlagen: Datenbanken* im WS15/16

Harald Lang, Linnea Passing (gdb@in.tum.de)
http://www-db.in.tum.de/teaching/ws1516/grundlagen/

Blatt Nr. 11

HINWEIS: Dieses Übungsblatt in der Woche vom 11.1. bis 15.1.2016 in den Tutorien besprochen.

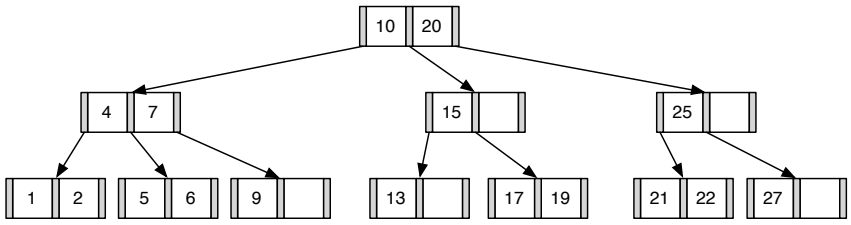
Hausaufgabe 1

Es sollen alle ca. 10 Milliarden Menschen in einer erweiterbaren Hashtabelle verwaltet werden. In jede Seite passen ca. 200 Einträge, durchschnittlich sind die Seiten halb voll. Je Verweis werden 4 Byte benötigt, da die Musterlösung aus einer Zeit stammt, in der es defakto nur Maschinen mit 32 bit CPU Architektur gab. Wie viel Speicherplatz verbraucht das Verzeichnis mindestens?

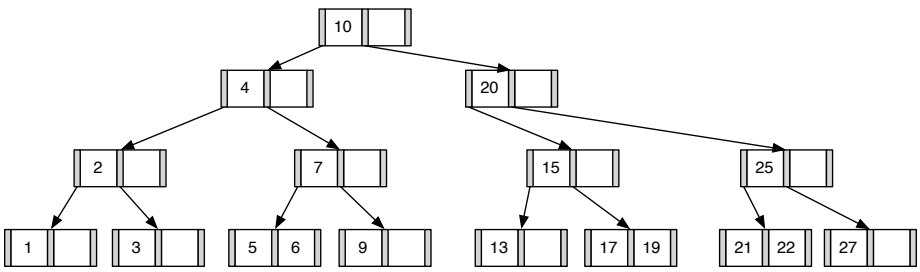
Lösung:

Das Verzeichnis enthält die Verweise auf alle Seiten (= Buckets), in dem die Einträge gehalten werden. Da pro Seite durchschnittlich 100 Einträge Platz haben, benötigen wir insgesamt $10^{10}/100 = 10^8$ Seiten. Um 10^8 Seiten zu referenzieren benötigen wir mindestens $\log_2 10^8$ Bits. Da dies eine positive ganze Zahl sein muss, ist die Anzahl der benötigten Bits $\lceil \log_2 10^8 \rceil$. Hiermit können $2^{\lceil \log_2 10^8 \rceil}$ Verweise im Verzeichnis abgelegt werden, da die Anzahl der Verweise in einem Verzeichnis immer einer 2er-Potenz entspricht. Pro Verweis werden 4 Byte benötigt, so dass das Verzeichnis eine Größe von $2^{\lceil \log_2 10^8 \rceil} \cdot 4$ Byte, also ungefähr 512 MB hat.

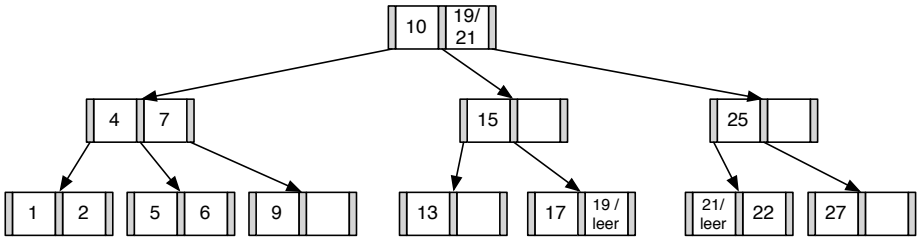
Hausaufgabe 2



1. Fügen Sie die 3 in den gezeigten B-Baum ein. Zeichnen Sie das Endergebnis. Zeichnen Sie jeweils den kompletten Baum oder machen Sie **deutlich**, falls Teile des Baumes unverändert bleiben. Verwenden Sie den aus der Vorlesung bekannten Algorithmus.
   Das Ergebnis sieht wie folgt aus:



2. Entfernen Sie aus dem **ursprünglichen Baum** den Eintrag 20. Zeichnen Sie das Ergebnis der Operation. Sollte es mehrere richtige Lösungen geben, genügt es, wenn Sie hier eine angeben. Zeichnen Sie jeweils den kompletten Baum oder machen Sie **deutlich**, falls Teile des Baumes unverändert bleiben. Verwenden Sie den aus der Vorlesung bekannten Algorithmus.
   Das Ergebnis sollte wie folgt aussehen:

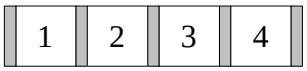


Hausaufgabe 3

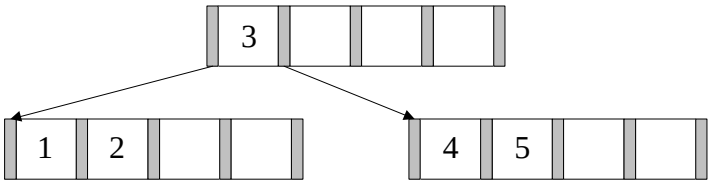
Fügen Sie in einen anfänglich leeren B-Baum mit $k = 2$ die Zahlen eins bis zwanzig in aufsteigender Reihenfolge ein. Was fällt Ihnen dabei auf?

Lösung:

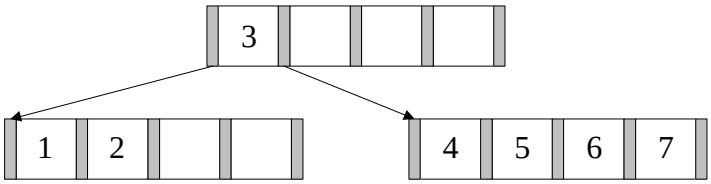
Nachdem man die Zahlen 1 bis 4 eingefügt hat, liegt folgender B-Baum vor:



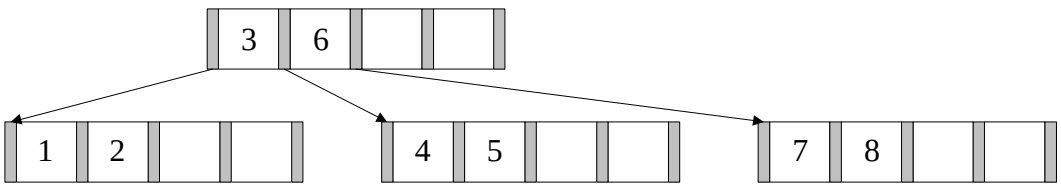
Beim Einfügen von 5 wird der Knoten gespalten und man erhält eine neue Wurzel.



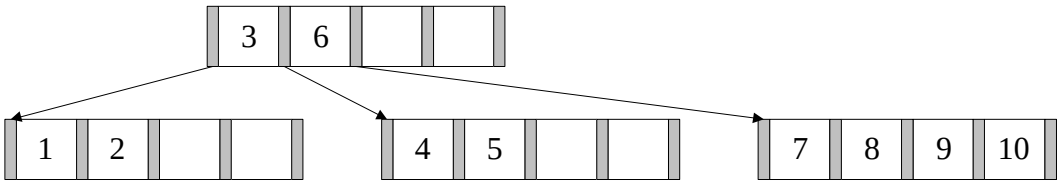
Die nächsten beiden Zahlen lassen sich wieder ohne Probleme einfügen.



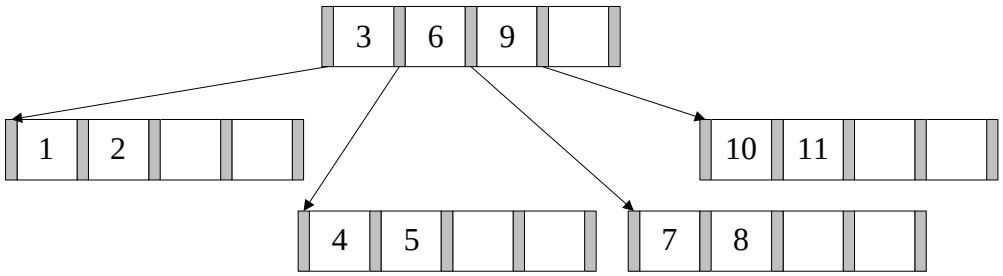
Beim Einfügen der 8 kommt es erneut zum Überlauf. Die 6 wandert in die Wurzel.



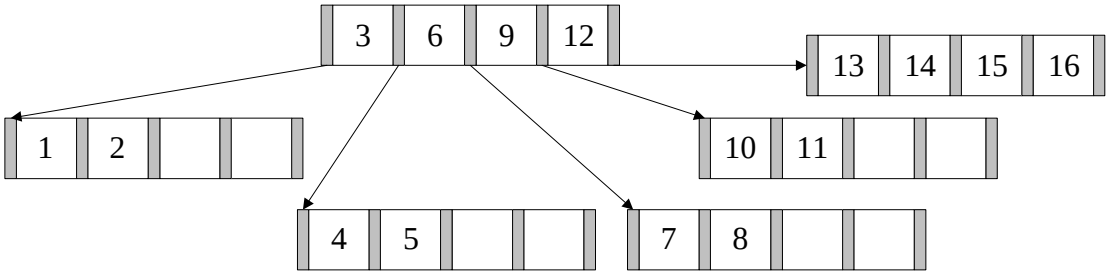
9 und 10 lassen sich wieder ohne Probleme einfügen. Bei 11 kommt es zum Überlauf.



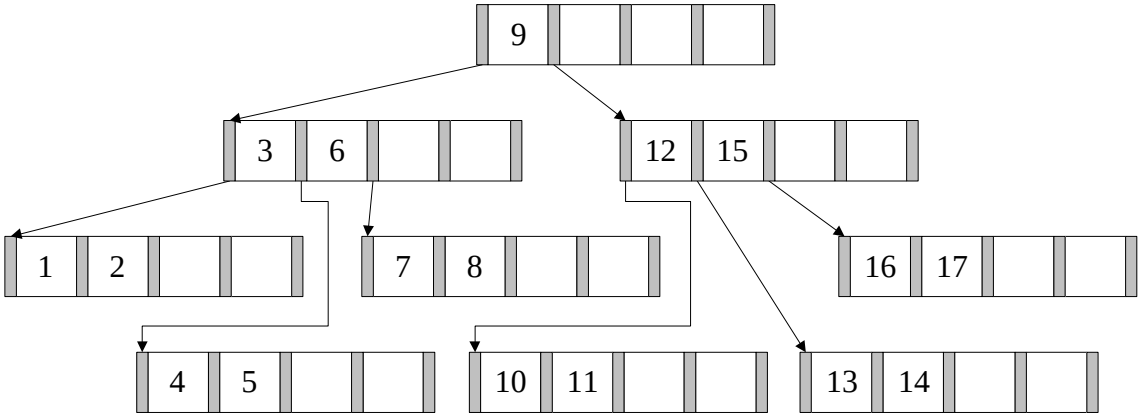
Nach dem Aufspalten erhält man dann:



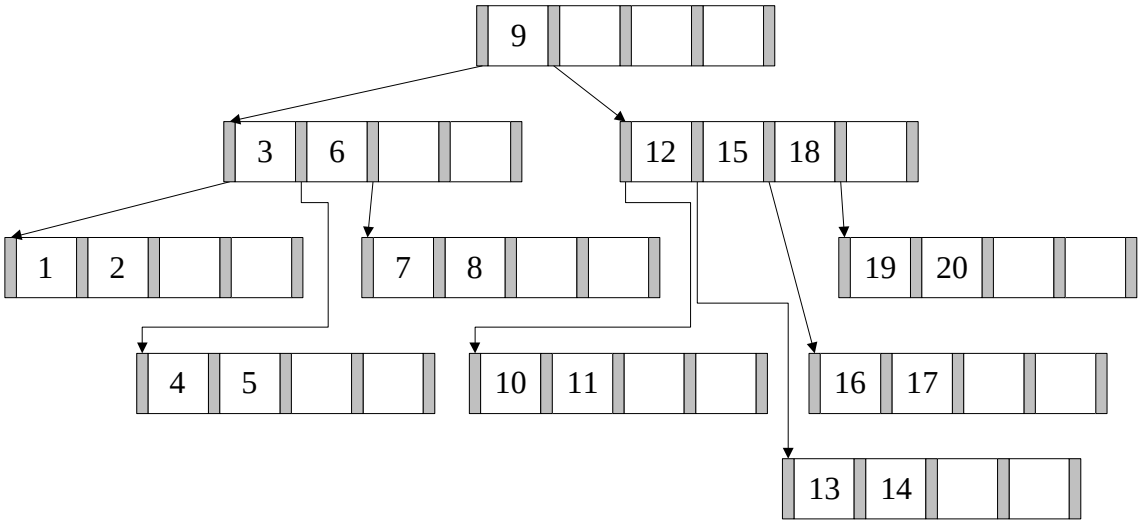
Es werden nun die nächsten Zahlen bis 16 analog eingefügt.



Bei 17 kommt es dann wieder zum Überlauf.



Fügt man nun noch die restlichen Zahlen ein, erhält man folgenden B-Baum:



Es fällt auf, dass der B-Baum nahezu minimale Auslastung aufweist. Dies liegt daran, dass eine aufsteigende Zahlenfolge sequentiell in den Baum eingefügt wird. Nach dem Aufspalten einer Seite in zwei Seiten werden dann in die Seite, die die kleineren Datensätze enthält, keine weiteren Werte mehr eingefügt. Allgemein ist das sortierte Einfügen der Schlüssel in einen B-Baum eine sehr schlechte Idee, da dies zu einer sehr geringen Auslastung führt.

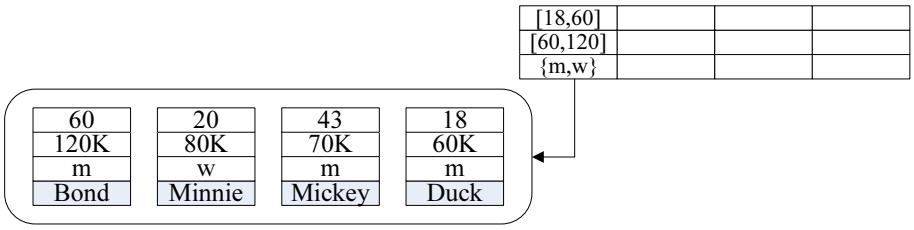
Hausaufgabe 4

Indizieren Sie die folgende Tabelle in einem R-Baum mit den Dimensionen *Alter*, *Gehalt* und *Geschlecht*. Nehmen Sie an, dass (i) die Kapazität der inneren Knoten sowie der Blattknoten gleich 4 ist und (ii) die Tabelle die Einfügereihenfolge festlegt. Illustrieren Sie die einzelnen Phasen im Aufbau des R-Baums.

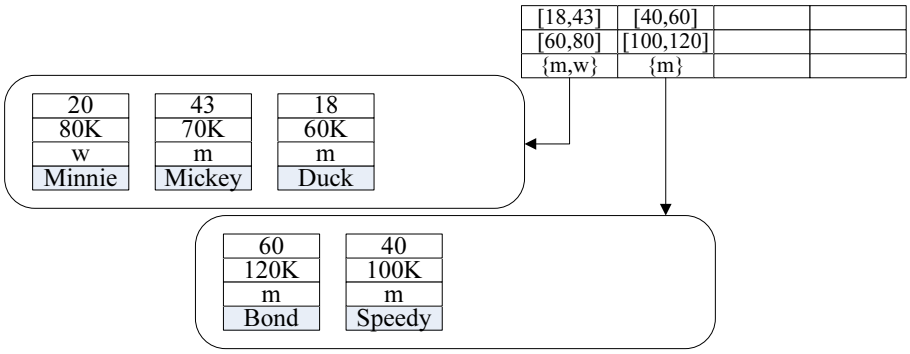
Name	**Alter**	**Gehalt**	**Geschlecht**
Bond	60	120 K	männlich
Minnie	20	80 K	weiblich
Mickey	43	70 K	männlich
Duck	18	60 K	männlich
Speedy	40	100 K	männlich
Bert	45	55 K	männlich
Ernie	41	45 K	männlich
Urmel	35	112 K	neutral
Bill	25	110 K	männlich
Lucie	65	95 K	weiblich
Jan	41	60 K	männlich
Sepp	50	65 K	männlich

Lösung:

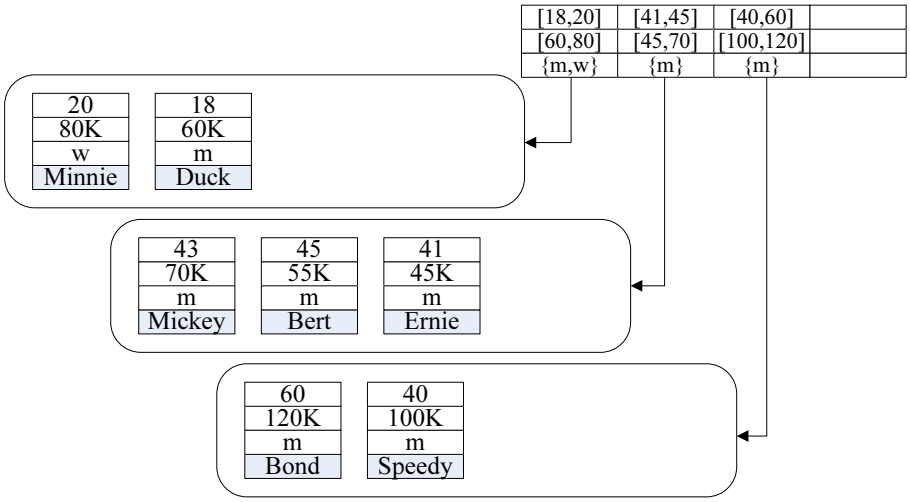
Gemäß Aufgabenstellung habe die Blätter des R-Baums eine Maximalbelegung von 4. Damit können die Datensätze für Bond, Minnie, Mickey und Duck eingefügt werden, ohne dass eine Überlaufbehandlung erforderlich ist.



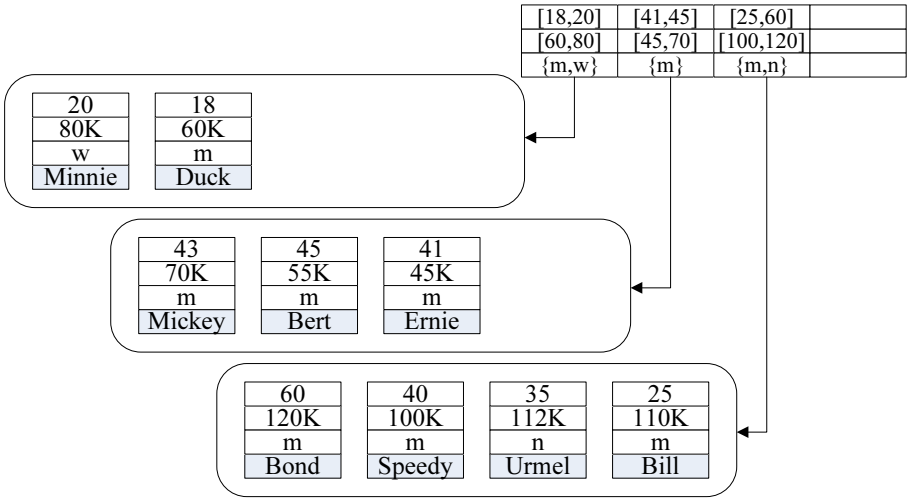
Fügt man zusätzlich Speedy in den R-Baum ein, so muss das Blatt aufgespalten werden.



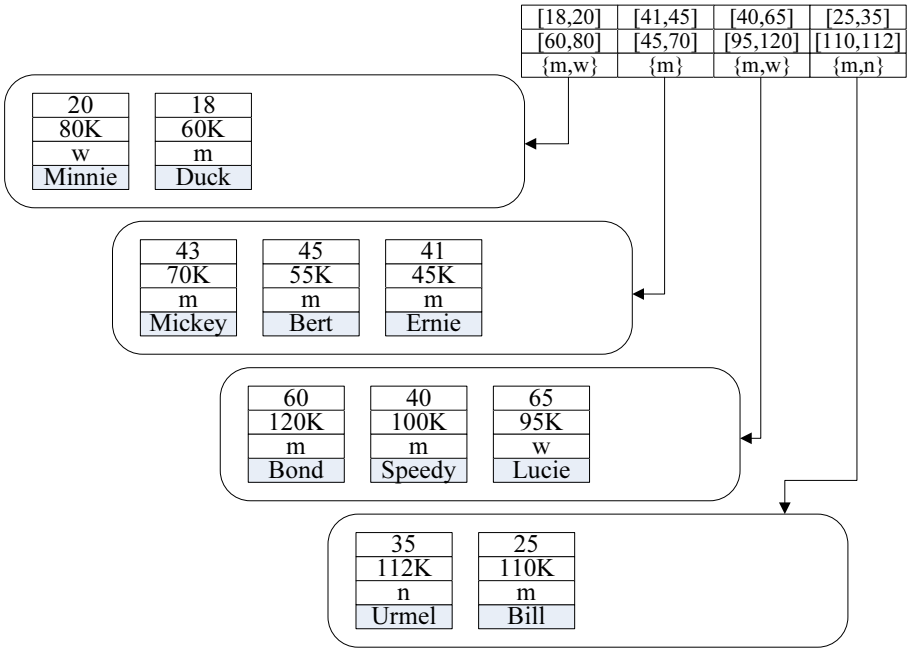
Die Datensätze für Bert und Ernie müssen bezüglich ihrer Ausprägungen in das erste Blatt des R-Baums eingefügt werden. Fügt man zuerst Bert und dann Ernie in den R-Baum ein, so muss erneut eine Überlaufbehandlung, d.h. ein Aufspalten des Blattes erfolgen.



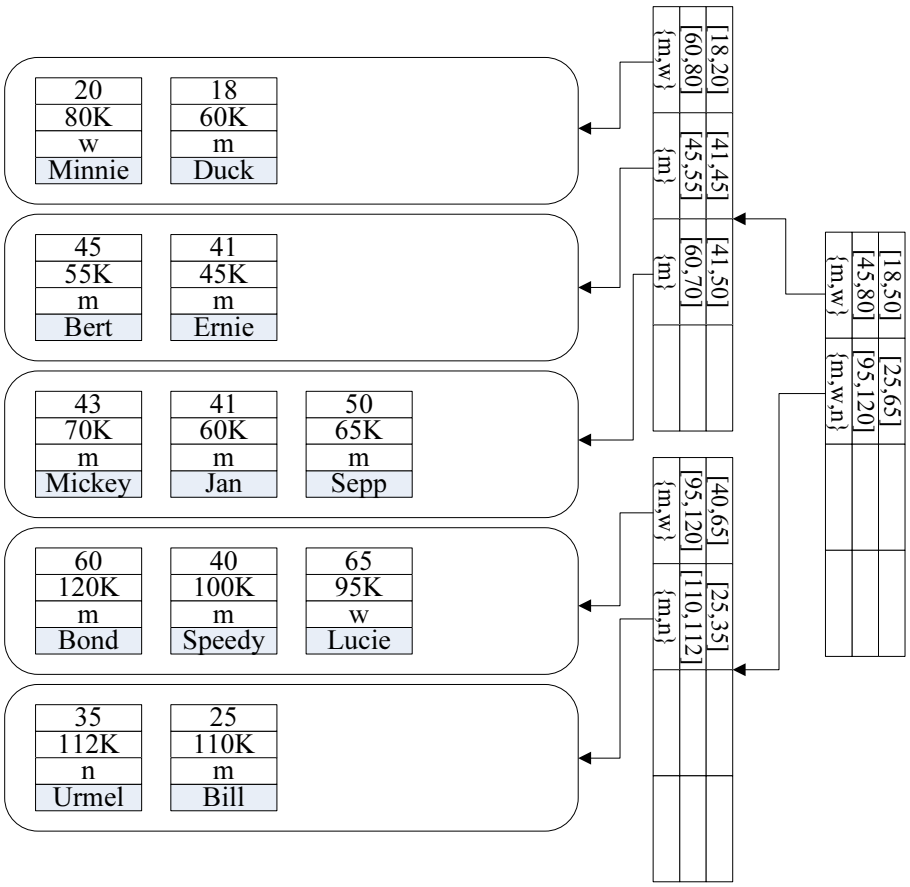
Die folgenden Datensätze für Urmel und Bill passen wieder in den bisherigen R-Baum, ohne eine Aufspaltung hervorzurufen.



Ein Aufspalten ist erst beim Einfügen von Lucie notwendig. Wir gruppieren hier Bill und Urmel in einen Behälter und sehen einen zweiten Behälter für Speedy, Lucie und Bond vor. Nimmt man als Gütekriterium das Volumen, das durch die jeweiligen Behälter beschrieben wird, so hat diese Aufteilung Vorteile.



Anschließend werden noch die Datensätze für Jan und Sepp eingefügt. Beide würden am besten in das Blatt passen, das bereits von Mickey, Bert und Ernie belegt ist. Das Blatt ist damit Überbelegt und es erfolgt eine Spaltung. Diese Spaltung propagiert sich bis zum Vaterknoten, der zugleich der Wurzelknoten ist. Deshalb muss eine neue Wurzel erstellt werden. Folgende Darstellung zeigt den finalen R-Baum für unsere Beispielausprägung. Abbildung 1 stellt die Datensätze dreidimensional dar.



Hausaufgabe 5

Bestimmen Sie alle Kandidatenschlüssel der Relation R. Wenden Sie den Dekompositionsalgorithmus an, um die Relation R in die BCNF zu zerlegen und unterstreichen Sie die Schlüssel der Teilrelationen des Endergebnisses.

$$R = \{A, B, C, D, E, F\}$$

FDs:

1. $B \rightarrow DA$
2. $DEF \rightarrow B$
3. $C \rightarrow EA$

Prüfen Sie als erstes FD 1) ob Sie für die Zerlegung geeignet ist und - falls dies der Fall ist - verwenden Sie diese im ersten Zerlegungsschritt. Für diese Aufgabe ist zu bedenken, dass die oben angegebenen FDs eine Charakterisierung der insgesamt geltenden FDs sind. Die Menge der geltenden FDs ist größer. Wieso? Wie muss dies beim Dekompositionsalgorithmus genutzt werden?

Lösung:

- Dekompositionsalgorithmus:
  - Starte mit $Z := \{R\}$.

- R in BCNF? - Nein, $B \to DA$ verletzt die BCNF.
  - Zerlegung anhand FD $B \to DA$, da $\{B\}$ kein Superschlüssel:
    $R_1 = \{A, B, D\}$ mit den FDs $F_1 = \{B \to DA\}$,
    $R_2 = \{B, C, E, F\}$ mit den FDs $F_2 = \{C \to E\}$, FD (2) geht verloren und FD (3) geht "teilweise" verloren: wenn $C \to AE$ gilt, dann gilt auch $C \to A$ und $C \to E$ (Dekompositionsregel), aber lediglich $C \to E$ bleibt erhalten.
    $Z := \{R_1, R_2\}$
- R_1 in BCNF? - Ja.
- R_2 in BCNF? - Nein, $C \to E$ verletzt die BCNF.
  - Zerlegung anhand FD $C \to E$, da $\{C\}$ kein Superschlüssel:
    $R_{2.1} = \{C, E\}$ mit den FDs $F_{2.1} = \{C \to E\}$,
    $R_{2.2} = \{B, C, F\}$ mit ausschließlich trivialen FDs.
    $Z := \{R_1, R_{2.1}, R_{2.2}\}$
- $R_{2.1}$ in BCNF? - Ja.
- $R_{2.2}$ in BCNF? - Ja.

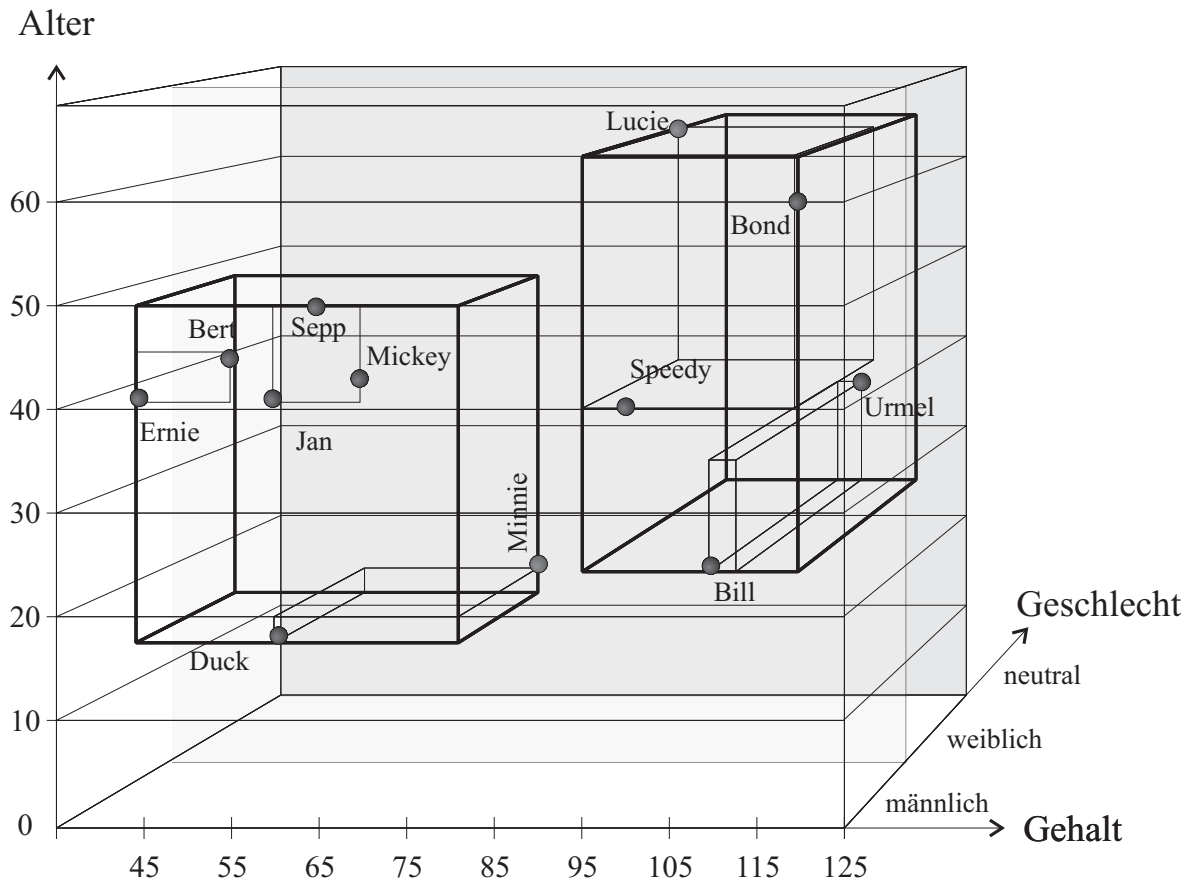


Abbildung 1: Der R-Baum mit drei Dimensionen (Gehalt, Alter, Geschlecht)

- Ergebnis:

$$\begin{aligned} R_1 &= \{A, \underline{B}, D\} \\ R_{2.1} &= \{\underline{C}, E\} \\ R_{2.2} &= \{\underline{B}, \underline{C}, \underline{F}\} \end{aligned}$$

Im Allgemeinen ist eine gegebene FD-Menge weder minimal noch vollständig. Die angegebenen FDs enthalten also möglicherweise Redundanzen einerseits und andererseits werden triviale Abhängigkeiten i.d.R. nicht explizit mit angegeben. Bei der Ausführung des Dekompositionsalgorithmus müssen jedoch alle *geltenden* FDs betrachtet werden, die sich mit Hilfe der Axiome von Armstrong herleiten lassen (F^+). So gilt im obigen Beispiel in R_2 die FD $C \rightarrow E$, obwohl diese nicht explizit angegeben war.