Datenbanken
Wintersemester 2020/21
Prof. Dr. W. May

2. Übungsblatt: Algebra

Besprechung voraussichtlich am 9./10./16./17.12.2020

Aufgabe 1 (Relationale Anfragen an Mondial: Bedingungen) Geben Sie Ausdrücke der relationalen Algebra für die folgenden Anfragen an die Mondial-Datenbank an:

a) Die Namen aller Städte, die mehr als 1.000.000 Einwohner haben.
b) Die Namen aller Städte, die mehr Einwohner als Neuseeland haben.
c) Die Namen aller Städte, in denen mehr als 25% der Bevölkerung des jeweiligen Landes leben.

Für spätere Übungsblätter:

- Geben Sie dieselben Anfragen in SQL an.

---

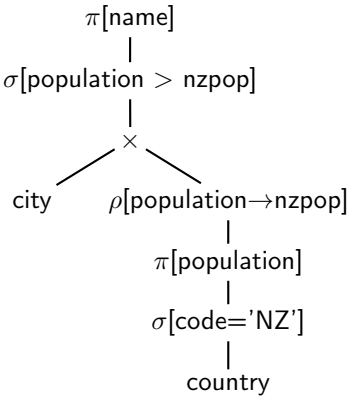
a) Einfache Selektion:

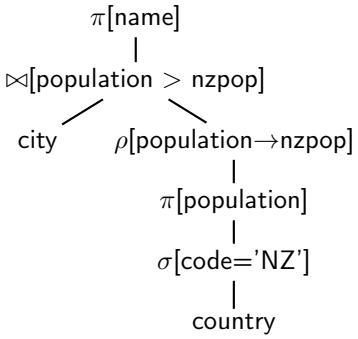
$$\pi[\mathsf{name}](\sigma[\mathsf{population} > 1000000](\mathsf{city}))$$

b) Die Bevölkerungszahl von Neuseeland bekommt man als

$$\pi[\mathsf{population}](\sigma[\mathsf{code}=\text{'NZ'}](\mathsf{country})\ .$$

In der Selektionsbedingung in Teil (a) ist aber nur eine Konstante erlaubt. Hier kann nicht stattdessen eine Subquery (bzw. ihr Ergebnis) stehen.
Man benötigt also ein Join, in dem man jedes Land mit dem Ergebnis der Subquery joint und danach den Vergleich durchführt (bzw. kann diese beiden Schritte auch in einem Theta-(Semi-)Join zusammenfassen)





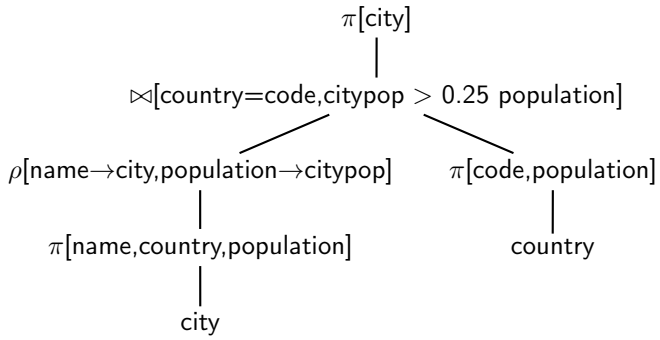
Hinweis: in SQL kann man die Subquery in die WHERE-Klausel einbauen, wenn deren Format einspaltig ist, und die Anfrage zur Laufzeit nur eine einzige Zeile ergibt:

```
SELECT name
FROM city
WHERE population > (SELECT population FROM country WHERE code='NZ')
```

oder als Join formulieren:

```
SELECT name
FROM city,  (SELECT population as nzpop FROM country WHERE code='NZ')
WHERE population > nzpop
```

c) Theta-Join (Bedingung: keine gleichnamigen Attribute!) mit Zusatzbedingung:



Aufgabe 2 (Äquivalenz von Ausdrücken) Gegeben seien folgende Relationen:

- R(A,B,C)
- S(A,E,F)
- T(A,H)

Die Wertebereiche aller nicht namensgleichen Attribute seien voneinander verschieden. Gegeben sei nun folgender relationaler Ausdruck:

$$\pi[E,H](\sigma[B=10]((R \bowtie T) \bowtie S))$$

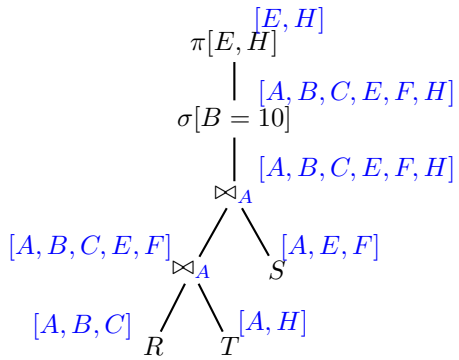
Sind die folgenden Ausdrücke äquivalent zu obigem Ausdruck? Begründen Sie Ihre Antwort.

a) $\pi[E,H]((\sigma[B=10](R)) \bowtie (\pi[A,E](S) \bowtie T))$

b) $\pi[E,H](\sigma[B=10](((\pi[B](R)) \bowtie (\pi[A,E](S))) \bowtie (\pi[A,H](T))))$

c) $\pi[E,H]((\pi[A,B](\sigma[B=10](R))) \bowtie ((\pi[A](S)) \bowtie T))$

The first expression as an algebra tree (annotated with the formats of the expressions and the natural join attributes)



a) ist äquivalent:
Der ursprüngliche Ausdruck:

$$\pi[E,H](\sigma[B=10]((R \bowtie T) \bowtie S))$$

Join ist assoziativ:

$$\pi[E,H](\sigma[B=10](R \bowtie (T \bowtie S)))$$

Attribut B existiert nur in Relation R, daher kann die Selektion vor dem Join ausgeführt werden:

$$\pi[E,H]((\sigma[B=10](R)) \bowtie (T \bowtie S))$$

Von S werden nur die Attribute A (im Join mit T) und E (in der abschließenden Projektion) benötigt, man kann die Projektion also auch gleich auf S ausführen:

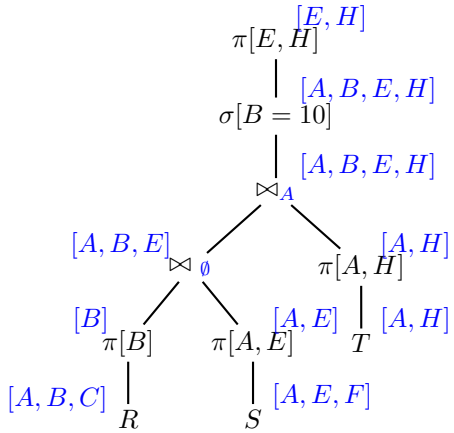
$$\pi[E,H]((\sigma[B=10](R)) \bowtie (T \bowtie (\pi[A,E]S)))$$

Join ist auch kommutativ:

$$\pi[E,H]((\sigma[B=10](R)) \bowtie ((\pi[A,E]S) \bowtie T))$$

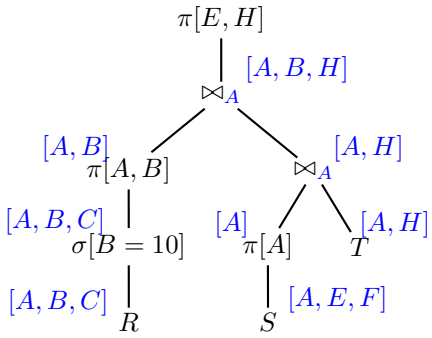
Der letzte Ausdruck ist der Ausdruck aus (a) und damit ist gezeigt dass die Äquivalenz gilt.
Hinweis: in diesem Ausdruck werden alle π und σ so früh wie möglich ausgeführt.

b) nicht äquivalent:



$\pi[B](R)$ eliminiert das für den Join mit S und T wichtige Attribut A, hier wird stattdessen im unteren Join das Kreuzprodukt ausgeführt.
$\pi[A,E](S)$ ist wie in (1) gezeigt zulässig – F wird nicht gebraucht.
$\pi[A,H](T)$ ist nutzlos (Identität), stört aber keinen.
Der Ausdruck ist syntaktisch in Ordnung, hat aber ein anderes Ergebnis (eine Obermenge, da ein Join-Kriterium wegfällt).
Dazu gibt man dann ein (kleines) Beispiel an, z.B. einen Datenbankzustand $\mathcal{S}$ mit $\mathcal{S}(R) = \{(a_0,b,c)\}$, $\mathcal{S}(S) = \{(a_1,e,f)\}$, $\mathcal{S}(T) = \{(a_1,h)\}$.
Die Auswertung des ersten Ausdrucks, $Q_0(\mathcal{S}) = \emptyset$, während die Auswertung des Ausdrucks in (a), $Q_a(\mathcal{S}) = \{(e,h)\}$ ergibt.

c) dieser Ausdruck ist syntaktisch garnicht zulässig:



$\pi[A](S)$ eliminiert das Attribut E, welches in der Ausgabe enthalten sein soll. Der Ausdruck ist damit syntaktisch nicht zulässig.

Aufgabe 3 (Relationale Anfragen an Mondial: Schweizer Sprachen) Geben Sie Ausdrücke der relationalen Algebra für die folgenden Anfragen an die Mondial-Datenbank an:

a) Alle Landescodes von Ländern, in denen eine Sprache gesprochen wird, die auch in der Schweiz gesprochen wird.

b) Alle Landescodes von Ländern, in denen ausschliesslich Sprachen gesprochen werden, die in der Schweiz nicht gesprochen werden.

c) Alle Landescodes von Ländern, in denen nur Sprachen gesprochen werden, die auch in der Schweiz gesprochen werden.

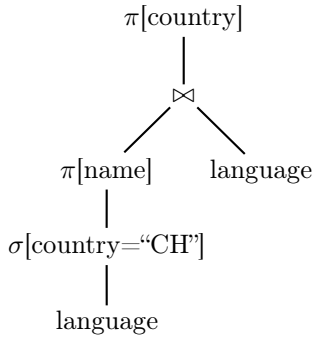
d) Alle Landescodes von Ländern, in denen alle Sprachen gesprochen werden, die in der Schweiz gesprochen werden.

Für spätere Übungsblätter:

- Geben Sie dieselben Anfragen in SQL an.

---

a) Join zur Formulierung einer "Auswahlbedingung". Verwende Tabelle language(country,name,percent), z.B. (CH, "german", 65).

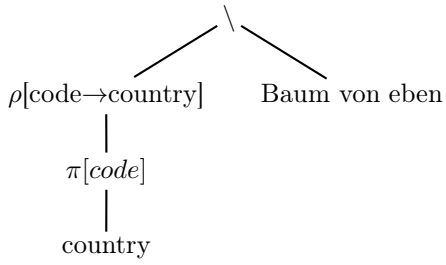


Der linke Ast ergibt die Tabelle

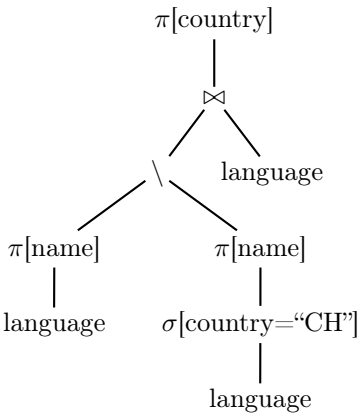
language
name
german
french
italian
romansch

Diese wird mit language gejoint (Join-Attribut "name"), womit genau die Einträge aus language übrigbleiben, die eine der aufgeführten Sprachen enthalten.

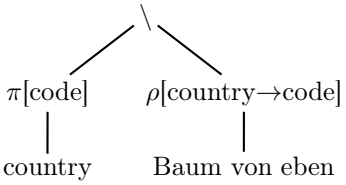
b) Genau alle Länder, die eben nicht aufgezählt wurden:



c) Löst man in mehreren Schritten. Erst alle Sprachen bestimmen, die nicht in CH gesprochen werden. Das Ergebnis wird mit language gejoint, und übrig bleiben die Einträge, die nicht-schweizer Sprachen beschreiben. Die Länder (bzw. Landescodes), in denen irgendeine nicht-schweizer Sprache gesprochen wird, bekommt man also folgendermassen:

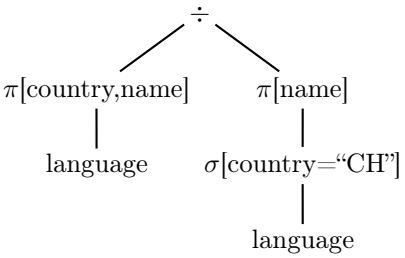


Bleibt nun, die Komplementmenge, eben die Länder in denen nur schweizer Sprachen gesprochen werden, zu bestimmen:



... schönes Beispiel dafür, wie man Subqueries in der WHERE-Klausel in der Algebra durch Joins realisiert.

d) "alle" – was wieder mit Division gelöst wird. $\pi[country, name](language)$ durch die Menge der schweizer Sprachen dividieren. Übrig bleiben die Länder die mit allen diesen Sprachen genannt werden:



Aufgabe 4 (Korrektheit der Join-Algorithmen) Die Definition des relationalen Join-Operators, $\bowtie$, auf den Folien ist nicht konstruktiv, sondern beschreibt "nur" *deklarativ*, aus welchen Tupeln das Ergebnis bestehen soll.

Beweisen Sie, dass die beiden folgenden Join-Algorithmen (a) und (b) korrekt sind, d.h. diese Menge von Tupeln liefern.

Gegeben seien Relationen $R(\bar{X})$ und $S(\bar{Y})$; berechnet werden soll $R \bowtie S$.

a) Nested-Loop-Join:

let $\bar{Z} := \bar{X} \cap \bar{Y}$;
let $T := \emptyset$;
for $r \in R$ do

```
    begin
      for s ∈ S do
        if r[Z̄] = s[Z̄] then
          begin
            μ := ein neues Tupel so dass μ(x) = r(x) für alle x ∈ X̄ und μ(y) = s(y) für alle y ∈ Ȳ;
            T := T ∪ {μ};
          end;
    end;
return T;
```

b) Loop-Index-Join: Hierbei ist ein Baumindex über den Join-Attributen $\bar{Z} := \bar{X} \cap \bar{Y}$ von S in der Datenbank vorhanden.
Nehmen Sie an, dass $\bar{Z}$ Schlüssel von S ist (dann erstellt ein Datenbanksystem diesen Index automatisch). Nehmen Sie weiter an, dass $\bar{Z}$ nur ein Attribut enthält, welches numerisch ist, und stellen Sie sich als Index einen Binärbaum wie (hoffentlich) in Info I besprochen vor. In den jeweiligen Baumknoten ist nicht nur der Wert, sondern auch eine Referenz auf das Tupel, das diesen Wert hat, enthalten.

```
let Z̄ := X̄ ∩ Ȳ;
let B := der Baumindex über S.Z̄;
let T := ∅;
for r ∈ R do
    begin
      if Suche nach r[Z̄] in B erfolgreich
        begin
          s = das vom Baumknoten referenzierte Tupel in S;
          μ := ein neues Tupel so dass μ(x) = r(x) für alle x ∈ X̄ und μ(y) = s(y) für alle y ∈ Ȳ;
          T := T ∪ {μ};
        end;
    end;
return T;
```

c) Wie ist die Komplexität der beiden Algorithmen?

---

a) Nested Loop. Die äußere Schleife iteriert über alle Tupel $r \in R$, in der inneren Schleife wird jedes Tupel $s \in S$ getestet, ob es "passt" (d.h. $r[\bar{Z}] = s[\bar{Z}]$). Wenn ja, wird das kombinierte Tupel rs zum Ergebnis dazugenommen.
R enthält m Tupel, S enthält n Tupel, geschachtelte Schleife, also Komplexität $O(n \cdot m)$.
Korrektheitsbeweis: Zu zeigen ist

$$A_{lgo} := \{\mu \in \mathrm{Tup}(\overline{XY}) : \text{Ergebnis des Algorithmus enthält } \mu\} = \\ \{\mu \in \mathrm{Tup}(\overline{XY}) \mid \mu[\bar{X}] \in R \text{ und } \mu[\bar{Y}] \in S\} = D_{ef} \quad \text{of } R \bowtie S$$

Zu zeigen sind zwei Richtungen: "$\subseteq$" (Korrektheit: alle ausgegebenen Ergebnisse sind richtig), und "$\supseteq$" (Vollständigkeit: der Algo findet alle Ergebnisse). Üblicherweise ist die Korrektheit einfacher zu zeigen (da sie die Entwicklung des Algorithmus widerspiegelt), als die Vollständigkeit (diese ist aber wichtiger, da man evtl. etwas übersehen hat).

"$A_{lgo} \subseteq D_{ef}$ " Wenn μ ein Ergebnistupel ist, dann gibt es Tupel $r \in R$ und $s \in S$ so dass $r[\bar{Z}] = s[\bar{Z}]$ ist (die in dem "if"-Statement an diese Variablen gebundenen Tupel), nach denen μ so erzeugt wird, dass $\mu(x) = r(x)$ für alle $x \in \bar{X}$ und $\mu(y) = s(y)$ für alle $y \in \bar{Y} \setminus \bar{X}$, also $\mu = rs$ gilt.
Offensichtlich ist dann genau $r = \mu[\bar{X}]$ und $s = \mu[\bar{Y}]$, und damit $\mu \in R \bowtie S$.

"$D_{ef} \subseteq A_{lgo}$" Wenn $\mu \in R \bowtie S$, ist also $r' := \mu[\bar{X}] \in R$ und $s' := \mu[\bar{Y}] \in S$. Die Variable r der äußeren Schleife enthält irgendwann dieses Tupel r', und im inneren Durchlauf enthält die Variable s irgendwann das Tupel s'. Da $Z = \bar{X} \cup \bar{Y}$ ist $r'[\bar{Z}] = s'[\bar{Z}]$, womit die "if"-Bedingung zu "wahr" ausgewertet wird, und das entsprechende Ergebnistupel erzeugt wird.

b) Loop-Index-Join. Die äußere Schleife iteriert über alle Tupel $r \in R$. Für jedes r wird über den Index auf alle Tupel $s \in S$ zugegriffen, für die $s[\bar{Z}] = r[\bar{Z}]$ ist, und das kombinierte Tupel rs zum Ergebnis dazugenommen.

Komplexität: Komplexität: n Durchläufe der äußeren Schleife. Der Index-Lookup benötigt $O(\log n)$ (für ausgeglichene Binärbäume $O(\log_2 n)$, für die in Datenbanken üblichen B- bzw. B*-Bäume (siehe späteres Kapitel der Vorlesung) z.B. $O(\log_{200} n)$).

Hat man exakt ein Tupel pro Indexeintrag zu verarbeiten, ist die Komplexität somit $O(n \log m)$.

Sind es mehrere (keine bis viele) Tupel (beim allgemeinen Join, wenn $\bar{Z}$ nicht Schlüssel von S ist) ist immerhin garantiert, dass pro Zugriff ein Ergebnis geliefert wird. Wenn man die Anzahl der Ergebnisse nicht (anhand Wissens über die Anwendung) abschätzen kann, kann man nur $O(\max(n \log m, \#\text{Ergebnisse}))$ als Komplexitätsabschätzung angeben.

Beweis: Der Beweis funktioniert wie oben:

"$A_{lgo} \subseteq D_{ef}$ " wie oben. Man könnte auch einfach argumentieren, dass das Ergebnis eine Teilmenge dessen aus (a) ist.

"$D_{ef} \subseteq A_{lgo}$" Diese Richtung ist etwas interessanter. Wenn $\mu \in R \bowtie S$, ist also $r' := \mu[\bar{X}] \in R$ und $s' := \mu[\bar{Y}] \in S$. Die Variable r der äußeren Schleife enthält irgendwann dieses Tupel r'. Da $Z = \bar{X} \cap \bar{Y}$ ist, ist $s'[\bar{Z}] = r'[\bar{Z}]$, womit der Indexeintrag zu diesem Wert $r'[\bar{Z}]$ auf s' verweisen muss (im Prinzip reduziert man das Problem auf die Vollständigkeit des Indexes). Damit wird für r' und s' das entsprechende Ergebnistupel erzeugt.

c) siehe (a) und (b).

Aufgabe 5 (Division mit Basisoperationen) Beweisen Sie, daß die in der Vorlesung angebene Darstellung der Division durch relationale Basisoperatoren als

$$r \div s = \pi[\bar{Z}](r) - \pi[\bar{Z}]((\pi[\bar{Z}](r) \bowtie s) - r)$$

mit $r \in \mathsf{Rel}(\bar{X})$, $s \in \mathsf{Rel}(\bar{Y})$ und $\bar{Z} = \bar{X} \setminus \bar{Y}$ äquivalent zu der gegebenen Definition

$$r \div s = \{\mu \in \mathsf{Tup}(\bar{Z}) \mid \mu \in \pi[\bar{Z}](R) \wedge \{\mu\} \bowtie s \subseteq r\}$$

ist.

Veranschaulichen Sie sich Ihre Überlegungen anhand des Beispiels "Geben Sie die Namen derjenigen Organisationen an, die auf jedem Kontinent mindestens ein Mitglied haben".

Nach Definition der Division ist

$$r \div s = \{\mu \in \mathsf{Tup}(\bar{Z}) \mid \mu \in \pi[\bar{Z}](R) \wedge \{\mu\} \bowtie s \subseteq r\} \quad (*)$$

Betrachte

$$r \div s = \pi[\bar{Z}](r) - \underbrace{\pi[\bar{Z}]((\pi[\bar{Z}](r) \bowtie s) - r)}_{(***)} \quad (**)$$

Im Beispiel ist r die Menge aller Paare $(org, cont)$ so dass Organisation o ein Mitglied auf Kontinent $cont$ hat. s ist die Menge der (Namen der) Kontinente. $r \div s$ ist das gesuchte Ergebnis.

Im allgemeinen Fall (z.B. "... auf allen Kontinenten mit mehr als 10.000.000 km^2 Fläche") kann r auch Tupel mit $\bar{Y}$-Werten (Kontinenten) enthalten, die in $\{\mu\} \bowtie s$ nicht gefordert werden. Die Division prüft nur, ob die geforderten Kombinationen vorhanden sind.

Vorwärtsbeweis

(1) $(**) \subseteq (*)$: Korrektheit des Ausdrucks. Alle Ergebnisse erfüllen die Charakterisierung:
Sei $\mu \in (**)$.
μ ist also in $\pi[\bar{Z}](r)$, also $\in \mathsf{Tup}(\bar{Z})$ – soweit nicht besonders interessant.
Interessanter ist der zweite Teil, der aussagt, dass μ nicht in $\pi[\bar{Z}]((\pi[\bar{Z}](r) \bowtie s) - r)$ ist. μ ist also kein $\bar{Z}$-Tupel (im Beispiel: keine Organisation), so dass bei $(\{\mu\} \bowtie s) - r$ ein Rest übrigbleibt. D.h., $\{\mu\} \bowtie s \subseteq r$, und damit ist μ ein Tupel, das $(*)$ erfüllt.

(2) $(*) \subseteq (**)$: Vollständigkeit der Charakterisierung – Ergebnis enthält alle geforderten Werte:
Sei μ ein Tupel über $\bar{Z}$ (Beispiel: eine Organisation), $\mu \in (*)$, also $\mu \in \pi[\bar{Z}](R)$ und $\{\mu\} \bowtie s \subseteq r$.
Damit also schon mal $\mu \in \pi[\bar{Z}](r)$; jetzt also noch zu zeigen, dass μ nicht in $\mu \notin (***)$ ist.
Aus $\{\mu\} \bowtie s \subseteq r$ folgt, dass $(\{\mu\} \bowtie s) - r$ leer ist. Damit ist auch μ nicht in $\pi[\bar{Z}]((t \bowtie s) - r$, für beliebige Mengen $t \subseteq \mathsf{Tup}\bar{Z}$, also nicht in $(***)$.
Hinweis: Man könnte anstelle $(**)$ auch

$$r \div s = \pi[\bar{Z}](r) - \underbrace{\pi[\bar{Z}]((\mathsf{Tup}(\bar{Z}) \bowtie s) - r)}_{(***)} \quad (*)$$

schreiben. Damit müsste man allerdings eine unendliche Menge von Werten überprüfen. Durch die Einschränkung auf $\pi[\bar{Z}](r)$ erhält man eine endliche Ausgangsmenge. (Im Beispiel entspräche $\mathsf{Tup}(\bar{Z})$ der Menge aller möglichen Zeichenketten als Namen von Organisationen, die nicht mit jedem Kontinent in r auftreten – natürlich kommen dabei als Antworten nur solche in Frage, die in $\pi[\bar{Z}](r)$ überhaupt enthalten sind).

Alternative: Widerspruchsbeweis (verwendet im Prinzip natürlich dieselbe Argumentation):

(1) $(**) \subseteq (*)$: Korrektheit des Ausdrucks. Alle Ergebnisse erfüllen die Charakterisierung:
Sei $\mu \in (**)$.
Annahme: $\mu \notin (*)$, d.h. $\{\mu\} \bowtie s \not\subseteq r$.
Dann gibt es also ein $\nu \in s$, so dass $\mu\nu \notin r$.
$\mu \in \pi[\bar{Z}](r)$ nach Konstruktion von $(**)$
Schauen wir uns wieder $(***)$ an: Da $\mu \in \pi[\bar{Z}](r)$ ist, ist $\mu\nu \in \pi[\bar{Z}](r) \bowtie s$ und ja $\notin r$, also $\mu\nu \in (\pi[\bar{Z}](r) \bowtie s) - r$ und damit $\mu = \pi[\bar{Z}](diesem) = (***)$.
Womit μ nicht in $(*)$ ist. Widerspruch zur Voraussetzung.

(2) $(*) \subseteq (**)$: Vollständigkeit der Charakterisierung – Ergebnis enthält alle geforderten Werte:
Sei $\mu \in (*)$, also $\{\mu\} \bowtie s \subseteq r$. Für alle $\nu \in s$ ist also $\mu\nu \in r$.
Damit also erstmal $\mu \in \pi[\bar{Z}](r)$.
Zu zeigen: $\mu \notin (***)$.
Annahme: $\mu \in (***)$ – also gäbe es ein ν', so dass $\mu\nu' \in (\pi[\bar{Z}](r) \bowtie s) - r$ ist. Der Anteil $\pi[\bar{Z}](\mu\nu')$ ist gerade μ, also müsste $\nu' \in s$ sein, damit dies erfüllt ist, und $\mu\nu' \notin r$ – im Widerspruch zu oben.
Also $\mu \in (**)$.

Aufgabe 6 (Definition der Division) Betrachten Sie (i) die Definition der relationalen Division aus der Vorlesung:

$$r \div s = \{\mu \in \mathrm{Tup}(\bar{Z}) \mid \mu \in \pi[\bar{Z}](r) \wedge \{\mu\} \times s \subseteq r\}$$

sowie (ii) die kürzere Definition

$$r \cdot\!|\!\cdot\, s = \{\mu \in \mathrm{Tup}(\bar{Z}) \mid \{\mu\} \times s \subseteq r\} \ .$$

Sind (i) und (ii) äquivalent, bzw. warum nicht? Wo liegt das Problem?

Die Definitionen (i) und (ii) beschreiben fast dasselbe, (ii) ist allerdings problematisch, wenn $s = \emptyset$ ist: dann ist die Bedingung $\{\mu\} \times s \subseteq r$ äquivalent zu $\emptyset \subseteq r$, was immer erfüllt ist, und damit $r \cdot|\cdot s = \text{Tup}(\bar{Z})$, also die Menge aller Tupel über $\bar{Z}$ – unklar mit welchen Werten.

Ob diese Menge "unendlich groß" ist, hängt dann von der Sichtweise ab: betrachtet man wirklich alle möglichen Werte (Zahlen, Zeichenketten, ...)? Aus logischer Sicht hängt das Ergebnis vom abstrakten zugrundeliegenden *Domain* ab (der endlich oder unendlich sein kann) und jede Obermenge der in der Datenbank vorkommenden Werte (des sog. "aktiven Domains") sein kann. Damit wäre das Ergebnis des Ausdrucks nicht nur von dem tatsächlichen Datenbankinhalt abhängig. Es würde alle $|\bar{Z}|$-Tupel über diesem (dem Anfragesteller unbekannten) Domain umfassen, und wäre damit nicht sinnvoll (siehe Abschnitt des Foliensatzes über *domain-independence* von Formeln des relationalen Kalküls).

Aufgabe 7 (Tupeloperatoren vs. Relationale Operatoren) In der Vorlesung wurden auf einzelnen Tupeln nur die Operatoren Projektion $\pi[\bar{X}](\mu)$, Selektion $\sigma[\alpha](\mu)$ und Renaming $\rho[A \to B](\mu)$ definiert. Die relationalen Operatoren wurden dann auf Basis dieser Operatoren definiert, wobei für das Join nur eine deklarative, auf tupelbasierter Projektion aufbauende Definition gegeben wurde.

- Geben Sie die Definition des relationalen Joins an: "Sei $r \in \text{Rel}(\bar{X})$ and $s \in \text{Rel}(\bar{Y})$. Dann ist $r \bowtie s = \{$*was gehört hier hin?*$\}$.
- Überlegen Sie, wie ein Join-Operator für Tupel $\mu \in \text{Tup}(\bar{X})$, $\nu \in \text{Tup}(\bar{Y})$, also $\mu \bowtie \nu$, definiert werden kann, und geben Sie darauf basierend eine Definition des relationalen Join-Operators an.
- Kann man eine entsprechende Definition auch für die Division angeben?

- $r \bowtie s = \{\mu \in \text{Tup}(\overline{XY}) \mid \mu[\bar{X}] \in r \text{ and } \mu[\bar{Y}] \in s\}$.
- Tupelbasiertes Join: sei $\mu \in \text{Tup}(\bar{X})$, $\nu \in \text{Tup}(\bar{Y})$. Die Verknüpfung $\mu \bowtie \nu$ bezeichnet das Join zweier Tupel und ist definiert als

$$\mu \bowtie \nu = \begin{cases} \mu \cup \nu \in \text{Tup}(\overline{XY}) & \text{falls } \mu[\bar{X} \cap \bar{Y}] = \nu[\bar{X} \cap \bar{Y}] \\ \text{nichts} & \text{sonst} \end{cases}$$

  Was ist dabei $\mu \cup \nu$? Dank des formalen Vorgehens ist das ganz einfach und klar: Jedes Tupel μ ist eine Abbildung, die (Spalten)namen auf Werte abbildet, Mit $\bar{X} = (X_1, \ldots, X_n)$ und $\bar{Y} = (Y_1, \ldots, Y_m)$ ist $\mu = \{X_1 \mapsto v_1, \ldots, X_n \mapsto v_n\}$ und $\nu = \{Y_1 \mapsto w_1, \ldots, Y_m \mapsto w_m\}$ und wobei $v_i \in \mathsf{dom}(X_i)$ und $w_i \in \mathsf{dom}(Y_i)$ ist. Also ist $\mu \cup \nu$ die Vereinigung dieser Einzelabbildungen, d.h.

$$(\mu \cup \nu)(A) = \begin{cases} \mu(A) & \text{falls } A \in \bar{X} \\ \nu(A) & \text{falls } A \in \bar{Y} \\ \text{nicht definiert} & \text{sonst} \end{cases}$$

  Da gefordert wird, dass μ und ν auf $\bar{X} \cap \bar{Y}$ übereinstimmen, ist $\mu \cup \nu$ wohldefiniert.
  Relationales Join damit für $r \in \text{Rel}(\bar{X})$ und $s \in \text{Rel}(\bar{Y})$: Ergebnisformat ist (natürlich wie in der originalen Definition) $\overline{XY}$.
  Ergebnisrelation:

$$r \bowtie s = \{\mu \bowtie \nu \mid \mu \in r \text{ and } \nu \in s\} \ .$$

  Damit hat man jetzt eine konstruktive Charakterisierung des Joins (die die Berechnung als nested-loop-join direkt nahelegt).

- Division: läßt sich nicht auf eine Tupeloperation reduzieren. Die Division ist in der grundlegenden Definition eine Mengenoperation.

Aufgabe 8 (Äquivalenzen: Join, Division, Differenz) Seien $R(\bar{X}), S(\bar{Y})$ Relations-Schemata. Zeigen oder widerlegen Sie:

(a) Sei $\bar{X} \cap \bar{Y} = \emptyset$.

$$(R \bowtie S) \div S \equiv R.$$

(b) Sei $\bar{X} = \bar{Y}$ und $\bar{Z} \subseteq \bar{X}$.

$$\pi[\bar{Z}](R - S) \equiv \pi[\bar{Z}]R - \pi[\bar{Z}]S.$$

(zu a) Nach Voraussetzung sind $R(\bar{X})$ und $S(\bar{Y})$ Relationen, die kein gemeinsames Attribut haben. Damit gilt $(\bar{X} \cup \bar{Y}) - \bar{Y} = \bar{X}$.

Behauptung: dann gilt: $(R \bowtie S) \div S = R$

Definition der Division: Sei $\bar{Y} \subset \bar{X}$, $\bar{Z} = \bar{X} - \bar{Y}$.

$$r \div s = \{\mu \in \mathsf{Tup}(\bar{Z}) \mid \{\mu\} \times s \subseteq r\} = \pi[\bar{Z}](r) - \pi[\bar{Z}]((\pi[\bar{Z}](r) \times s) - r).$$

Man kann entweder die Umschreibung der Division in Algebra-Operatoren, oder die Definition verwenden:

$$\begin{array}{ll}
(R \bowtie S) \div S & \text{keine gemeinsamen Attribute} \\
= (R \times S) \div S & \text{Umschreibung durch Algebra-Operatoren} \\
= \pi[\bar{X}](R \times S) - \pi[\bar{X}]((\pi[\bar{X}](R \times S) \times S) - R \times S) & \\
= R - \pi[\bar{X}]((R \times S) - (R \times S)) & \\
= R - \pi[\bar{X}](\emptyset) & \\
= R &
\end{array}$$

(dies ist die interessantere, aber längere Beweis), oder

$$\begin{array}{ll}
(R \bowtie S) \div S & \text{keine gemeinsamen Attribute} \\
= (R \times S) \div S & \text{Definition der Division: mit } R \times S \subseteq \mathsf{Tup}(\overline{XY}) \\
= \{\mu \in \mathsf{Tup}(\bar{X}) \mid \{\mu\} \times S \subseteq (R \times S)\} & \\
= \{\mu \in \mathsf{Tup}(\bar{X}) \mid \mu \in R\} = R &
\end{array}$$

(zu b) $\pi[\bar{Z}](R - S) \neq \pi[\bar{Z}](R) - \pi[\bar{Z}](S)$

Bsp:

Sei $\bar{X} = \bar{Y} = \{A, B\}$ und $\bar{Z} \subseteq \bar{X} = \{A\}$

R	
A	B
1	2

S	
A	B
1	3

$\pi[A](R - S) = \pi[A](R) = \{(A : 1)\}$ (da $R - S = R$ ist), und $\pi[A](R) - \pi[A](S) = \{(A : 1)\} - \{(A : 1)\} = \emptyset$

Aufgabe 9 (Algebra: Minimale- und Maximale Anzahl von Tupeln) Die Relationen $R(\bar{X})$ und $S(\bar{Y})$ enthalten n bzw. m Tupel. Wie groß ist die maximale und minimale Anzahl von Tupeln, die das Ergebnis folgender Operationen (bei geeigneten $\bar{X}, \bar{Y}$) enthalten kann?

a) $R \cup S$

b) $R \bowtie S$

c) $\sigma[C](R) \times S$, für eine Bedingung C

d) $\pi[\bar{Y}](R) - S$

e) $R \div S$

---

a) Nur zulässig wenn $\bar{X} = \bar{Y}$.
   max: $n + m$ (keine gemeinsamen Tupel)
   min: $max\{n, m\}$ (Teilmengenbeziehung)

b) max: $n * m$, wenn entweder $\bar{X} \cap \bar{Y} = \emptyset$ oder im natürlichen Join jedes Tupel R mit jedem Tupel aus S auf den gemeinsamen Attributen $\bar{X} \cap \bar{Y}$ zusammenpasst (d.h., diese Attribute haben in allen Tupeln beider Relationen den selben Wert).
   min: 0 (natürliches Join mit $\bar{X} \cap \bar{Y} \supsetneq \emptyset$, bei dem aber nichts übrigbleibt, da die Einträge der gemeinsamen Spalte(n) nicht zusammenpassen.
   Hinweis: man sieht, dass ein Join also die Ergebnismenge sowohl deutlich vergrößern, als auch einschränkend wirken kann.
   - Verhältnis zwischen $|R \bowtie S|$ und $|R|$ bezeichnet man als *Selektivität* des Joins (bzgl. R).
   - ist $\bar{X} \cap \bar{Y}$ z.B. Schlüssel von S und Fremdschlüssel in R, so hat man $|R \bowtie S| = |R|$.

c) max: $n * m$ (alle Tupel in R erfüllen C)
   min: 0 (kein Tupel in R erfüllt C)

d) Nur sinnvoll wenn $\bar{Y} \subseteq \bar{X}$.
   max: n, wenn $S = \emptyset$, oder zumindest $S \cap \pi[\bar{Y}](R) = \emptyset$, und bei der Projektion keine Duplikate (die wegfallen würden) entstehen.
   min: 0, wenn $\pi[\bar{Y}](R) \subseteq S$.

e) max: n/m – ganzzahlige Division, wenn alle Werte $\pi[\bar{X} \setminus \bar{Y}](R)$ mit jedem der Werte aus S in R vorkommen.
   Insbesondere, wenn $|S| = m = 1$ ist, und jedes Tupel aus R mit diesem Wert zusammen vorkommt, gilt $R \div S = \pi[\bar{X} \setminus \bar{Y}](R)$ und $|R \div S| = n$.
   Wenn noch extremer $|S| = \emptyset$ ist, ist $R \div S = \pi[\bar{X} \setminus \bar{Y}](R)$, und ebenfalls $|R \div S| = n$.
   min: 0, wenn kein Wert die durch die Division bzgl. S gestellte "für alle"-Bedingung erfüllt.

Aufgabe 10 (Transitive Hülle) Gegeben sei eine Relation $R(A, B)$. Skizzieren Sie einen Algorithmus, der, bestehend aus Operationen der relationalen Algebra und einer while-Schleife, die transitive Hülle der Relation R berechnet.

Hinweis: Die transitive Hülle einer Relation R, bezeichnet als R^*, ergibt sich wie folgt: betrachte z.B. eine Relation $R(von, nach)$ von Flugverbindungen. R^2 ist dann die Menge aller Verbindungen, die über eine Zwischenlandung zustandekommen, etc; R^n sind also diejenigen, Verbindungen, die sich aus n Teilverbindungen zusammensetzen. Die unendliche Vereinigung $R \cup R^2 \cup R^3 \cup ...$ für $R \to \infty$ wird dann als R^* bezeichnet. In einer endlichen Datenbasis benötigt man nur endlich viele Schritte um diese zu berechnen. Ein anderes beliebtes Beispiel ist die aus $Kind(x, y)$ berechnete Vorfahren-Relation.

Problem der Algebra und SQL: jedes R^n kann ausgedrückt werden - aber die beliebig große Vereinigung nicht. Man weiß nicht, wo man abbrechen soll.

Über $R(A, B)$ soll die transitive Hülle gebildet werden. T und S sind binäre Relationen über A und B.

$S := \emptyset$
$T := R$
while $T - S \neq \emptyset$ do
    begin
        $S := T$;
        $T := T \cup \pi[T.A, R.B](\sigma[T.B = R.A](T \times R))$;
    end.

In diesem Programm enthält S jeweils den Wert von T aus der letzten Iteration der Schleife. Die Berechnung endet, wenn S und T übereinstimmen, d.h. während der zuletzt ausgeführten Iteration wurden keine weiteren Tupel mehr zu T hinzugefügt.

```
% ?- river(N,R,L,S,LE,A,SLa,SLo,Mt,SE,ELa,ELo,EE).
% ?- river(N,R,L,S,LE,A,SLa,SLo,Mt,SE,ELa,ELo,EE), S \= null.
% ?- lake(N,A,De,El,T,R,La,Lo), R \= null.
toSea(N,S) :- river(N,_R,_L,S,_,_,_,_,_,_,_,_,_,_), S \= null.
toSea(N,S) :- river(N,R,_L,_S,_,_,_,_,_,_,_,_,_,_), R \= null, toSea(R,S).
toSeaLake(N,S) :- lake(N,_,_,_,_,R,_,_), R \= null, toSea(R,S).
toSea(N,S) :- river(N,_R,L,_,_,_,_,_,_,_,_,_,_,_), L \= null, toSeaLake(L,S).
```