Einführung in Datenbanken

Übung 12: GROUP BY, Relationale Algebra

Prof. Dr. Stefan Brass
PD Dr. Alexander Hinneburg
Martin-Luther-Universität Halle-Wittenberg

Wintersemester 2023/24

http://www.informatik.uni-halle.de/~brass/db23/

Inhalt

Haben Sie Fragen/Anmerkungen?

- Die Übung ist dafür da, dass Sie Fragen stellen können.
- Sie soll keineswegs ein Monolog werden.
- Haben Sie Fragen?
  - Organisatorisches?
  - Fragen zum Vorlesungsstoff?
- Haben Sie interessante neue Einsichten, die Sie teilen wollen?

Lehrevaluation (1)

- Sie müssten von Evaluationsbüro eine EMail bekommen haben, die einen Code bzw. Link enthält, mit dem Sie an einer Lehrevaluation dieser Vorlesung teilnehmen können.
- Diese Umfrage ist anonym: Ich bekomme nur eine Zusammenstellung der Ergebnise vom Evaluationsbüro, nicht die einzelnen Formulare.

  Es wird gewünscht, dass ich diese Ergebnisse dann in StudIP einstelle.

  Ich tue das üblicherweise, aber diese Daten dürfen sich keinesfalls über den Kreis der Teilnehmer hinaus weiter verbreiten.
- Daher kann ich auch bei irgendwie ungewöhnlichen Angaben zu Studienprogramm, Semester etc. nicht den Zusammenhang zu Ihren Antworten bei anderen Fragen herstellen.

  Ich erfahre allerdings vor Bewertung der Klausur Ihre Gesamtnote für mich (Das war in Pittsburgh bewusst anders.). Ich gehe professionell damit um.

Lehrevaluation (2)

- Bitte nehmen Sie sich die Zeit, an der Lehrevaluation teilzunehmen.

  Auch dann, wenn Sie alles in Ordnung fanden. Es ist ja auch bei Produktbewertungen in Onlineshops so, dass, wer sich geärgert hat, eher motiviert ist, eine Bewertung zu schreiben. Der Dozent freut sich aber natürlich auch über positive Bewertungen.

- Sie können davon ausgehen, dass ich die Ergebnisse interessiert lese (vor allem auch die Freitext-Kommentare zu Verbesserungsvorschlägen).

  Da keine Rückfragen möglich sind, sollten Sie besonders bei schlechten Noten möglichst klar aufschreiben, was Ihnen nicht gefallen hat, und wie man es besser machen könnte. Aber auch, wenn Sie etwas besonders gut fanden, kann es nicht schaden, den Dozenten darauf aufmerksam zu machen, damit er nicht gerade das ändert.

Lehrevaluation (3)

- In Halle entscheiden die Dozenten selbst, ob sie an der Lehrevaluation teilnehmen (aber in drei Jahren müssen $\geq$ 2 Veranstaltungen jedes Dozenten evaluiert werden).

  [Evaluations-Ordnung]

  Das führt zu besonders guten Durchschnittsnoten. Professoren, die sich gefühlt angestrengt haben, bekommen teils nur durchschnittliche Ergebnisse (natürlich war das dann kein Spitzenplatz). Es könnte auch ein Problem sein, dass die Statistik nur auf Fakultätsebene berechnet wird. Auch das war in Pittsburg anders: Jede Veranstaltung wurde evaluiert.

- Wenn Sie durch Teilnahme an der Lehrevaluation deutlich machen, dass Ihnen gute Lehre wichtig ist, wird das langfristig einen Beitrag zur Verbesserung der Lehre leisten.

  Letztes Jahr haben sich nur 20 Teilnehmer an der Umfrage beteiligt.
  In StudIP eingeschrieben waren 152 Teilnehmer. Den ersten Klausurtermin haben 81 Teilnehmer genutzt, den zweiten Klausurtermin 6.

Studiengangs-Evaluation

- Leider findet mehr oder weniger gleichzeitig eine Studiengangs-Evaluation statt, zumindest für die Studiengänge der Informatik/Bioinformatik.
- Da für Bioinformatik die Reakkreditierung ansteht, werden Bioinformatiker ganz besonders gebeten, das Formular für die Studiengangs-Evaluation auszufüllen.

  D.h. wenn Sie Bioinformatik studieren, und nur Zeit haben, ein Formular auszufüllen, sollten Sie sich für die Studiengangs-Evaluation entscheiden. Natürlich würde ich mich freuen, wenn Sie außerdem die Zeit finden, auch das Formular für meine Lehrveranstaltung auszufüllen.
- Das zeitliche Zusammentreffen ist wohl etwas unglücklich.

Inhalt

Beispiel-Datenbank

STUDENTEN

SID	VORNAME	NACHNAME	EMAIL
101	Lisa	Weiss	...
102	Michael	Grau	NULL
103	Daniel	Sommer	...
104	Iris	Winter	...

BEWERTUNGEN

SID	ATYP	ANR	PUNKTE
101	H	1	10
101	H	2	8
101	Z	1	12
102	H	1	9
102	H	2	9
102	Z	1	10
103	H	1	5
103	Z	1	7

AUFGABEN

ATYP	ANR	THEMA	MAXPT
H	1	ER	10
H	2	SQL	10
Z	1	SQL	14

Präsenzaufgabe: Erste RA-Anfrage (1)

Schreiben Sie folgende Anfrage in relationaler Algebra:

- Geben Sie die SID und die Punktzahl aller Bewertungen von Abgaben für Hausaufgabe 1 mit mindestens 8 Punkten aus.

  Schreiben Sie die Anfragen in ASCII, so dass RelaX sie versteht.
  [http://dbis-uibk.github.io/relax/calc/gist/8dc2652578ee12ae756a234c4cf21b3f]
  Auf die Webadresse wird auch unter „Übung“ auf der Webseite der Vorlesung verlinkt. Sie brauchen die URI nicht abzutippen.

Dies bezieht sich auf das bekannte Schema:

- STUDENTEN(SID, VORNAME, NACHNAME, EMAIL^o)
- AUFGABEN(ATYP, ANR, THEMA, MAXPT)
- BEWERTUNGEN(SID$\rightarrow$STUDENTEN, (ATYP, ANR)$\rightarrow$AUFGABEN, PUNKTE)

Präsenzaufgabe: Erste RA-Anfrage (2)

- Geben Sie die SID und die Punktzahl aller Bewertungen von Abgaben für Hausaufgabe 1 mit mindestens 8 Punkten aus.
- Lösung:

$$\pi_{\texttt{SID}, \texttt{PUNKTE}}\Big((\sigma_{\texttt{ATYP}=\texttt{'H'} \wedge \texttt{ANR} = 1 \wedge \texttt{PUNKTE} \geq 8}(\texttt{BEWERTUNGEN})\Big)$$

- ASCII-Schreibweise in RelaX:

```
pi SID, PUNKTE (
    sigma ATYP='H' and ANR=1 and PUNKTE >= 8 (
        BEWERTUNGEN))
```

- Ausgabe:

BEWERTUNGEN.SID	BEWERTUNGEN.PUNKTE
101	10
102	9

Präsenzaufgabe: Erste RA-Anfrage (3)

- Lösung (nochmals):

$$\pi_{\texttt{SID}, \texttt{PUNKTE}}\Big((\sigma_{\texttt{ATYP}=\texttt{'H'} \wedge \texttt{ANR} = 1 \wedge \texttt{PUNKTE} \geq 8}(\texttt{BEWERTUNGEN})\Big)$$

- Zum Vergleich:

```
SELECT SID, PUNKTE
FROM   BEWERTUNGEN
WHERE  ATYP = 'H' AND ANR = 1 AND PUNKTE >= 8
```

- Also:
  - Selektion σ entspricht der `WHERE`-Klausel.
  - Projektion π entspricht der `SELECT`-Klausel.

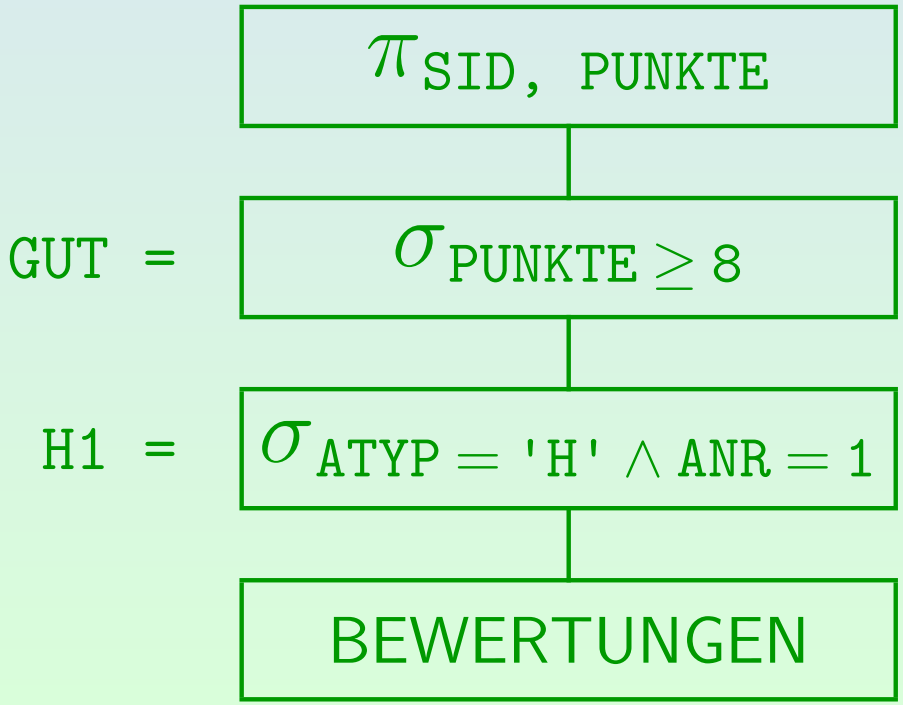
    Statt Projektion auf alle Spalten: Projektion ganz weglassen.

Präsenzaufgabe: Erste RA-Anfrage (4)

- In RelaX auch möglich mit Zwischenrelationen:

```
H1  = sigma ATYP = 'H' and ANR = 1 (BEWERTUNGEN)
GUT = sigma PUNKTE >= 8 (H1)
pi SID, PUNKTE (GUT)
```

Im Skript sind Zwischenrelationen/Sichten auch vorgesehen, dort wird „:=“ für die Zuweisung verwendet, und am Ende jeder Zuweisung steht „;“.



Inhalt

Hausaufgabe 11.1: HAVING (1)

- Geben Sie Name und Vorname von allen Komponisten aus, von denen es mindestens 5 Stücke in der Datenbank gibt.
- Geben Sie die Anzahl der Stücke mit aus, nennen Sie die Spalte `anz_stuecke` und sortieren Sie die Ausgabe danach absteigend (größte Anzahl zuerst).

  Stücke mit unbekanntem Komponisten (d.h. mit Nullwert in der Spalte `knr`) können Sie natürlich ignorieren. Sie können davon ausgehen, dass Name und Vorname zusammen einen Alternativschlüssel der Komponisten-Tabelle bilden.

name	vorname	anz_stuecke
Händel	Georg Friedrich	25
⋮	⋮	⋮
Beethoven	Ludwig van	5

10 Datensätze

Hausaufgabe 11.1: HAVING (2)

- KOMPONIST(KNR, NAME, VORNAME, GEBOREN, GESTORBEN°)
- STUECK(SNR, KNR→KOMPONIST, TITEL, TONART°, OPUS°)
- Lösung (Typisches Muster für Klausur-Aufgaben):

```
SELECT k.name, k.vorname, COUNT(*) anz_stuecke
FROM   komponist k, stueck s
WHERE  k.knr = s.knr
GROUP  BY k.knr, k.name, k.vorname
HAVING COUNT(*) >= 5
ORDER  BY anz_stuecke DESC
```

Stücke mit unbekanntem Komponisten werden durch den Join eliminiert.

Hausaufgabe 11.2: COUNT DISTINCT (1)

- Von welchen Orchestern gibt es Aufnahmen auf mindestens zwei CDs?
- Gemeint sind dabei die CD-Packs, die durch eine `CDNR` identifiziert werden (eine Doppel-CD hat nur eine `CDNR` und soll in dieser Rechnung auch nur als eine „CD“ zählen).
- Natürlich soll ein Nullwert als Orchester (z.B. bei Solo-Stücken) ausgeschlossen sein.
- Geben Sie den Namen des Orchesters und die Anzahl der CD-Packs aus.
- Sortieren Sie nach der Anzahl CD-Packs (größte zuerst) und innerhalb einer Anzahl alphabetisch nach dem Orchester-Namen.

Hausaufgabe 11.2: COUNT DISTINCT (2)

- Die erwartete Ausgabe ist:

orchester	anz_cds
Academy of St.Martin-in-the-Fields	3
London Festival Orchestra	3
Concentus musicus Wien	2
Die Zagreber Solisten	2
London Philharmonic Orchestra	2
Radio-Sinfonieorchester Stuttgart	2
Südwest-Studioorchester	2

7 Datensätze

Hausaufgabe 11.2: COUNT DISTINCT (3)

- Lösung:

```
SELECT orchester,
       COUNT(DISTINCT cdnr) AS anz_cds
FROM   aufnahme
WHERE  orchester IS NOT NULL
GROUP  BY orchester
HAVING COUNT(DISTINCT cdnr) >= 2
ORDER  BY anz_cds DESC, orchester
```

- Oft gibt es von einem Orchester mehrere Aufnahmen auf der gleichen CD. Deshalb ist hier `COUNT(DISTINCT ...)` wichtig.

Diese Aggregationsfunktion kommt auch in fast jeder Klausur vor.

Hausaufgabe 11.3: Geschachtelte Aggregation (1)

- Was ist die minimale, maximale und durchschnittliche Anzahl Stücke pro CD?

  Sie müssen also die Anzahl Stücke pro `cdnr` bestimmen, und jeweils durch `anz_cds` teilen (z.B. haben Doppel-CDs nur eine `cdnr`), und anschließend den Durchschnitt über diesen Anzahlen bilden. Runden Sie die Zahlen auf eine Nachkommestelle (geht mit der Funktion `ROUND` mit der Anzahl Nachkommastellen als zweitem Argument). Nennen Sie die Spalten „`Min`“, „`Max`“ und „`Durchschnitt`“ — auch mit dieser Groß-/Kleinschreibung.

- Die erwartete Antwort ist:

Min	Max	Durchschnitt
0.5	23.0	3.8

1 Datensatz

Hausaufgabe 11.3: Geschachtelte Aggregation (2)

- CD(CDNR, NAME, HERSTELLER, ANZ_CDS, GESAMTSPIELZEIT)
- AUFNAHME(CDNR→CD, SNR→STUECK, ORCHESTER°, LEITUNG°)
- Dies ist ein Beispiel für eine geschachtelte Aggregation:

```
WITH   stuecke_pro_cd AS
       (SELECT a.cdnr,
               COUNT(*)/c.anz_cds AS anz_stuecke
        FROM   aufnahme a, cd c
        WHERE  a.cdnr = c.cdnr
        GROUP  BY a.cdnr, c.anz_cds)
SELECT ROUND(MIN(anz_stuecke),1) AS "Min",
       ROUND(MAX(anz_stuecke),1) AS "Max",
       ROUND(AVG(anz_stuecke),1) AS "Durchschnitt"
FROM   stuecke_pro_cd
```

Hausaufgabe 11.4: Verschiedene Mengen (1)

- Geben Sie für jeden Komponisten, von dem mindestens drei Stücke mit bekannter Tonart (nicht Null) in der Datenbank stehen, aus,
  - wie viele Stücke es sind, und
  - was der Prozentsatz in Dur-Tonarten und
  - was der Prozentsatz in Moll-Tonarten ist.

  Den Komponisten identifizieren Sie durch Name und Vorname und ordnen die Ausgabe nach dem Namen, bei gleichem Namen nach dem Vornamen. Dur-Tonarten sind solche, die in „`-dur`“ enden, und Moll-Tonarten solche, die in „`-moll`“ enden. Bei dieser Aufgabe sollen nur Stücke mit definierter Tonart betrachtet werden (auch in den Prozent-Berechnungen). Geben Sie die Prozente gerundet ohne Nachkommastellen aus (nutzen Sie `ROUND`, und zwar die Variante mit nur einem Parameter). Denken Sie an mögliche Probleme mit der Integer-Division!

Hausaufgabe 11.4: Verschiedene Mengen (2)

- Die erwartete Antwort ist:

name	vorname	anz	prz_dur	prz_moll
Bach	Johann Sebastian	10	70	30
Händel	Georg Friedrich	21	62	38
Mozart	Leopold	6	100	0
Mozart	Wolfgang Amadeus	5	80	20
Prokofiev	Serge	4	75	25
Schubert	Franz	3	67	33
Telemann	Georg Philipp	15	87	13
Vivaldi	Antonio	8	50	50
Wolf-Ferrari	Ermanno	4	100	0

9 Datensätze

Die beiden letzten Spalten sollen eigentlich `prozent_dur` und `prozent_moll` heißen.

Hausaufgabe 11.4: Verschiedene Mengen (3)

```
SELECT k.name, k.vorname, COUNT(*) anz,
       ROUND(SUM(CASE WHEN s.tonart LIKE '%-dur'
                      THEN 1 ELSE 0 END
                 ) * 100.0 / COUNT(*)
             ) AS prozent_dur,
       ROUND(SUM(CASE WHEN s.tonart LIKE '%-moll'
                      THEN 1 ELSE 0 END
                 ) * 100.0 / COUNT(*)
             ) AS prozent_moll
FROM   komponist k, stueck s
WHERE  k.knr = s.knr AND s.tonart IS NOT NULL
GROUP  BY k.knr, k.name, k.vorname
HAVING COUNT(*) >= 3
ORDER  BY k.name, k.vorname
```

Hausaufgabe 11.4: Verschiedene Mengen (4)

- Statt `SUM(CASE ... THEN 1 ELSE 0 END)` geht auch:

```
COUNT(CASE WHEN s.tonart LIKE '%-dur'
THEN s.snr ELSE NULL END)
```

  Statt `s.snr` kann man irgendetwas schreiben, was nicht `NULL` ist, z.B. auch `1`.
  Tatsächlich kann man auch `ELSE NULL` weglassen (ist ohnehin Default).

- Man beachte, dass PostgreSQL die Integer-Division verwendet, wenn beide Argumente von einem Integer-Typ sind.
- Die Aggregationsfunktion `COUNT` liefert den Typ `bigint`.
- Z.B. funktioniert Folgendes nicht (liefert meist 0, sonst 100):

```
(COUNT(CASE ... END) / COUNT(*)) * 100
```

  Wenn jemand z.B. 4 Stücke in einer Dur-Tonart hat, und 5 Stücke insgesamt, berechnet PostgreSQL 4/5 als 0. Mal 100 bleibt 0.

Hausaufgabe 11.4: Verschiedene Mengen (5)

```
SELECT k.name, k.vorname, COUNT(*) anz,
       ROUND((SELECT COUNT(*) FROM stueck dur
              WHERE  dur.knr = k.knr
              AND    dur.tonart like '%-dur'
             )*100.0/COUNT(*)) AS prozent_dur,
       ROUND((SELECT COUNT(*) FROM stueck moll
              WHERE  moll.knr = k.knr
              AND    moll.tonart LIKE '%-moll'
             )*100.0/COUNT(*)) AS prozent_moll
FROM   komponist k, stueck s
WHERE  k.knr = s.knr and s.tonart is not null
GROUP  BY k.knr, k.name, k.vorname
HAVING COUNT(*) >= 3
ORDER  BY k.name, k.vorname
```

Hausaufgabe 11.4: Verschiedene Mengen (6)

```
WITH
gesamt AS (SELECT knr, count(*) AS anz
           FROM   stueck
           WHERE  TONART IS NOT NULL
           GROUP  BY knr),
dur AS     (SELECT knr, count(*) AS anz
           FROM   stueck
           WHERE  TONART LIKE '%-dur'
           GROUP  BY knr),
moll AS    (SELECT knr, count(*) AS anz
           FROM   stueck
           WHERE  TONART LIKE '%-moll'
           GROUP  BY knr)
SELECT     -- siehe nächste Folie
```

Hausaufgabe 11.4: Verschiedene Mengen (7)

```
SELECT k.name, k.vorname, g.anz,
       ROUND(COALESCE(d.anz,0) * 100.0 / g.anz)
             AS prozent_dur,
       ROUND(COALESCE(m.anz,0) * 100.0 / g.anz)
             AS prozent_moll
FROM   komponist k
       JOIN gesamt g    ON k.knr = g.knr
       LEFT JOIN dur d  ON k.knr = d.knr
       LEFT JOIN moll m ON k.knr = m.knr
WHERE  g.anz >= 3
ORDER  BY k.name, k.vorname
```

Der äußere Verbund („`LEFT JOIN`“) ist nötig, weil es auch Komponisten gibt, die keine Moll-Stücke haben (und es auch welche ohne Dur-Stücke geben könnte). Wenn ein Komponist in der Hilfstabelle `moll` nicht auftaucht, würde er bei einem normalen Verbund eliminiert. Der äußere Verbund wird unten noch geübt.

Hausaufgabe 11.4: Verschiedene Mengen (8)

- Wenn man den Outer Join vermeiden will, muss man bei Komponisten ohne Moll-Stücke einen 0-Eintrag erzeugen:

```
moll AS (SELECT knr, COUNT(*) AS anz
         FROM   stueck
         WHERE  tonart LIKE '%-moll'
         GROUP  BY knr
         UNION  ALL
         SELECT knr, 0 AS anz
         FROM   komponist k
         WHERE  NOT EXISTS
                (SELECT *
                 FROM   stueck s
                 WHERE  s.knr = k.knr
                 AND    s.tonart LIKE '%-moll'))
```

Hausaufgabe 11.4: Verschiedene Mengen (9)

- Auch mit einer Unteranfrage unter SELECT kann man die Zahl 0 für Komponisten ohne Moll-Stück erzeugen:

```
moll AS (SELECT k.knr,
                (SELECT COUNT(*)
                 FROM   stueck s
                 WHERE  s.knr = k.knr
                 AND    s.tonart LIKE '%-moll')
                 AS anz
         FROM   komponist k)
```

Obwohl im Beispielzustand zufällig alle Komponisten, bei denen es überhaupt Stücke mit definierter Tonart haben, auch mindestens ein Dur-Stück haben, kann man sich darauf natürlich nicht verlassen. D.h. auch bei den Dur-Anzahlen muss man wie hier dafür sorgen, dass ggf. auch die Anzahl 0 erzeugt wird.

Hausaufgabe 11.4: Verschiedene Mengen (10)

- War diese Hausaufgabe zu schwer?
  - A. Viel zu schwer. Frustrierend.
  - B. Herausfordernd, ich bin daran gewachsen.
  - C. Ok.
  - D. War doch einfach.

Hausaufgabe 11.5: Aggregation+Selbstverbund (1)

- Erstellen Sie eine Liste aller vorkommenden Instrumente in alphabetischer Reihenfolge mit einer fortlaufenden ID (beginnend bei 1).
- Sie können die ID berechnen, indem Sie zählen, wie viele Instrumente in der alphabetischen Reihenfolge vor dem aktuellen Instrument kommen (oder gleich sind).
- Sortieren Sie die Ausgabe nach der so berechneten ID.

Hausaufgabe 11.5: Aggregation+Selbstverbund (2)

- Die erwartete Antwort ist:

ID	Instrument
1	Bagpipes
2	Bariton
3	Bass
4	Cello
5	Drehleier
6	English Horn
7	Erzaehler
8	Horn
⋮	⋮
15	Violine

15 Datensätze

Hausaufgabe 11.5: Aggregation+Selbstverbund (3)

- Lösung:

```
SELECT COUNT(DISTINCT a.instrument) AS "ID",
       b.instrument as "Instrument"
FROM   solist a, solist b
WHERE  a.instrument <= b.instrument
GROUP  BY b.instrument
ORDER  BY "ID"
```

Sind Sie reif für die Klausur?

- Genauer: Für die SQL-Aufgaben in der Klausur.
- Wie lang haben Sie pro SQL-Anfrage gebraucht (bis eventuell auf Aufgabe 4)?
  - A. Eher weniger als 10 Minuten.
  - B. Zwischen 10 und 15 Minuten.
  - C. Teilweise deutlich länger als 15 Minuten.
- Denken Sie auch daran, dass Sie in der Klausur kein ChatGPT und kein Internet haben, und natürlich kein Kopieren vom Kollegen erlaubt ist.
- Sie dürfen aber 5 Blätter/10 Seiten mit Notizen mitbringen.

  Beim zweiten Termin nur 3 Blätter/6 Seiten.

Inhalt

Aufgabe

- Welche der folgenden Ausdrücke der relationalen Algebra sind syntaktisch korrekt?
  Was bedeuten sie?

  - ☐ $\texttt{STUDENTEN}$.
  - ☐ $\sigma_{\texttt{MAXPT} \neq 10}(\texttt{AUFGABEN})$.
  - ☐ $\pi_{\texttt{VORNAME}}(\pi_{\texttt{NACHNAME}}(\texttt{STUDENTEN}))$.
  - ☐ $\sigma_{\texttt{PUNKTE} \leq 5}(\sigma_{\texttt{PUNKTE} \geq 1}(\texttt{BEWERTUNGEN}))$.
  - ☐ $\sigma_{\texttt{PUNKTE}}(\pi_{\texttt{PUNKTE} = 10}(\texttt{BEWERTUNGEN}))$.

Kartesisches Produkt und Spalten-Namen (1)

- Natürlich muss man auch mehrere Relationen verknüpfen können.
- Die Basis-Operation dazu ist das kartesische Produkt $\times$.
- Das kartesische Produkt entspricht der `FROM`-Klausel:
  Es betrachtet sämtliche Kombinationen von Tupeln aus den beiden Relationen und „klebt“ die beiden Tupel aus jedem Paar zu einem Tupel zusammen.

  Die Tupelkalkül-Sichtweise auf die `FROM`-Klausel ist etwas anders: Dort werden Variablenbelegungen betrachtet, die die Variablen jeweils auf Tupel abbilden. Dort findet das „Zusammenkleben“ nicht statt, aber es werden auch alle Kombinationen von Tupeln betrachtet.
- Wenn R die Attribute A und B hat, und S die Attribute C und D, so hat $R \times S$ die Attribute A, B, C, D.

Kartesisches Produkt und Spalten-Namen (2)

- Beispiel:

A	B
1	2
3	4

$\times$

C	D
6	7
8	9

$=$

A	B	C	D
1	2	6	7
1	2	8	9
3	4	6	7
3	4	8	9

- Da Attributnamen innerhalb eines Tupels eindeutig sein müssen, kann man das kartesische Produkt nur anwenden, wenn R und S keine gemeinsamen Attribute haben.

  Da man Spalten umbenennen kann, ist das keine echte Einschränkung.

- Das kartesische Produkt `STUDENTEN` $\times$ `BEWERTUNGEN` wäre also ein Syntaxfehler, weil es zwei Spalten mit Namen `SID` geben müsste.

Kartesisches Produkt und Spalten-Namen (3)

- RelaX akzeptiert dagegen `STUDENTEN x BEWERTUNGEN`.

  [http://dbis-uibk.github.io/relax/calc/gist/8dc2652578ee12ae756a234c4cf21b3f]

- Dort besteht der Name $R.A$ jeder Spalte aus zwei Teilen:
  - Dem Relationen-Präfix R und
  - dem eigentlichen Spalten-Name A.
- Falls der eigentliche Spalten-Name eindeutig ist, reicht der, um die Spalte zu benennen.

  Das ist so ähnlich wie bei SQL, wo der Präfix eine Tupelvariable ist.

- Bei `STUDENTEN x BEWERTUNGEN` muss man die `SID`-Spalten also mit `STUDENTEN.SID` und `BEWERTUNGEN.SID` ansprechen.
- Dagegen hört die `PUNKTE`-Spalte sowohl auf den Namen „`PUNKTE`“ wie auf den Namen „`BEWERTUNGEN.PUNKTE`“.

Kartesisches Produkt und Spalten-Namen (4)

- Im Prinzip versteht RelaX also z.B.

```
sigma STUDENTEN.SID = BEWERTUNGEN.SID
      (STUDENTEN x BEWERTUNGEN)
```

- Damit gibt es jetzt aber das Problem, dass das nicht mit dem Skript kompatibel wäre.

  Sie müssen damit rechnen, dass es einen kleinen Punktabzug gibt, wenn Sie in Klausur oder Hausaufgaben schreiben, die zwar in RelaX funktionieren, aber nach dem Skript illegal wären.

- Glücklicherweise gibt es eine Lösung, die in RelaX funktioniert UND mit dem Skript kompatibel ist:

```
sigma S.SID = B.SID
      (rho S (STUDENTEN) x rho B (BEWERTUNGEN))
```

  Mit dem rho-Operator kann man einen Präfix explizit einführen. Wenn Sie das so machen, sind sie auf der sicheren Seite.

Kartesisches Produkt und Spalten-Namen (5)

- Nach dem Skript hat jede Spalte nur einen Namen. Wenn Sie mit dem ρ_X-Operator einen Präfix X für alle Attribute eingeführt haben, müssen Sie den auch immer schreiben.
- Die Projektion erlaubt bei Bedarf auch einzelne Spalten-Umbenennungen, auch in RelaX:

```
pi NAME <- NACHNAME, EMAIL (STUDENTEN)
```

- Auch Berechnungen sind in der Projektion möglich (wie im Skript):

```
pi NAME <- concat(VORNAME, concat(' ', NACHNAME))
   (STUDENTEN)
```

- RelaX nimmt bei $\cup$ die Spalten-Namen der linken Tabelle. Damit gilt in RelaX nicht: $R \cup S = S \cup R$.

  Nach den Definitionen im Skript gilt es dagegen.

Joins/Verbunde

- Wenn der natürliche Verbund funktioniert, erspart er Ihnen die Umbenennung:

```
pi VORNAME, NACHNAME, PUNKTE
    (sigma ATYP = 'H' and ANR = 1
        (STUDENTEN join BEWERTUNGEN))
```

Er funktioniert, wenn die Join-Spalten gleiche Namen haben und die einzigen gleich benannten Spalten sind.

- Ein Join mit expliziter Bedingung ist aber auch möglich:

```
pi S.VORNAME, S.NACHNAME, B.PUNKTE
    (sigma B.ATYP = 'H' and B.ANR = 1
        ((rho S STUDENTEN) join S.SID = B.SID
            (rho B BEWERTUNGEN)))
```

Um mit dem Skript kompatibel zu sein, muss man jetzt z.B. auch `S.VORNAME` schreiben. Bei RelaX würde `VORNAME` reichen.

Mengendifferenz

- Die Mengendifferenz kann man in RelaX „`-`“ oder „`\`“ oder „`except`“ schreiben.

  Weitere Mengenoperationen: `union` (∪) und `intersect` (∩).

- Z.B. Studenten ohne Hausaufgaben:

```
STUDENTEN join
    (pi SID (STUDENTEN) -
        pi SID (sigma ATYP = 'H' (BEWERTUNGEN)))
```

  Der natürliche Verbund dient hier nur als Filter: Es werden die Studenten ausgewählt, deren `SID` im Ergebnis der Mengendifferenz vorkommt.

- RelaX hat auch den im Skript erwähnten „Anti-Join“ als Abkürzung (nur das Symbol ist anders):

```
STUDENTEN anti join
    pi SID (sigma ATYP = 'H' (BEWERTUNGEN))
```

Inhalt

Joins in SQL (1)

- Geben Sie Vorname und Nachname von allen Studierenden aus, die Hausaufgabe 1 bearbeitet haben.
- Klassische Lösung:

```
SELECT S.VORNAME, S.NACHNAME
FROM   STUDENTEN S, BEWERTUNGEN B
WHERE  S.SID = B.SID
AND    B.ATYP = 'H' AND B.ANR = 1
```

- Lösung mit explizitem Join (mit `ON`):

```
SELECT S.VORNAME, S.NACHNAME
FROM   STUDENTEN S JOIN BEWERTUNGEN B
       ON S.SID = B.SID
WHERE  B.ATYP = 'H' AND B.ANR = 1
```

Joins in SQL (2)

- Lösung mit explizitem Join (mit `USING`):

```
SELECT S.VORNAME, S.NACHNAME
FROM   STUDENTEN S JOIN BEWERTUNGEN B
       USING (SID)
WHERE  B.ATYP = 'H' AND B.ANR = 1
```

- Lösung mit explizitem Join (mit `NATURAL`):

```
SELECT S.VORNAME, S.NACHNAME
FROM   STUDENTEN S NATURAL JOIN BEWERTUNGEN B
WHERE  B.ATYP = 'H' AND B.ANR = 1
```

- In beiden Fällen werden die Spalten `S.SID` und `B.SID` verschmolzen, man kann nicht mehr auf sie zugreifen, sondern nur auf `SID`.

Joins in SQL (3)

- Bei PostgreSQL und MySQL funktioniert Folgendes, nach dem Standard darf man nur `SID` schreiben:

```
SELECT S.SID, S.VORNAME, S.NACHNAME
FROM   STUDENTEN S NATURAL JOIN BEWERTUNGEN B
WHERE  B.ATYP = 'H' AND B.ANR = 1
```

- Man darf in PostgreSQL und MySQL (MariaDB 5.5.68) aber auch einfach `SID` schreiben (standard-konform).

  Auch `B.SID` wird akzeptiert.

- Bei Join mit `ON` wäre es dagegen ein Fehler, nur `SID` zu schreiben (`column reference "sid" is ambiguous`).

Joins in SQL (4)

- Bei `NATURAL JOIN` ist das Risiko
  - dass es Spalten gibt, an die man nicht gedacht hat, die zufällig auch in beiden Tabellen vorkommen, die aber nicht als Verbundspalten gedacht sind.
  - dass es durch später hinzugefügte Spalten zu einer Veränderung der Join-Bedingung kommt.
- Daher empfehlen wir, `USING` zu verwenden (oder `ON`).

  `ON` war zumindest in der Übergangszeit deutlich portabler als `USING`.
- In der relationalen Algebra dürfen Sie dagegen den natürlichen Verbund gern einsetzen.

  Dort gibt es die Variante mit `USING` nicht, und die Anfragen sind nur Übungsaufgaben, stehen also nicht in Programmen, die Jahre/Jahrzehnte benutzt werden.

Äußerer Verbund in SQL (1)

- Test-Daten:

R	
X	Y
1	4
2	4
3	5

S	
Y	Z
4	10
5	11
6	12

- Was liefert folgende Anfrage?

```
SELECT * FROM R LEFT JOIN S USING (Y)
```

Option A		
X	Y	Z
1	4	10
2	4	10
3	5	11

Option B		
Y	X	Z
4	1	10
4	2	10
5	3	11

Option C		
X	Y	Z
1	4	10
2	4	10
3	5	11
NULL	6	12

Äußerer Verbund in SQL (2)

- Test-Daten:

R	
X	Y
1	4
2	4
3	5

S	
Y	Z
4	10
5	11
6	12

- Was liefert folgende Anfrage?

```
SELECT * FROM R RIGHT JOIN S USING (Y)
```

Option A		
Y	X	Z
4	1	10
4	2	10
5	3	11

Option B		
Y	X	Z
4	1	10
4	2	10
5	3	11
6	NULL	12

Äußerer Verbund in SQL (3)

- Test-Daten:

R	
X	Y
1	4
2	4
3	5

S	
Y	Z
4	10
5	11
6	12

- Was liefert folgende Anfrage?

```
SELECT * FROM R RIGHT JOIN S USING (Y)
    WHERE X < 3
```

Option A		
Y	X	Z
4	1	10
4	2	10

Option B		
Y	X	Z
4	1	10
4	2	10
6	NULL	12

Option C		
Y	X	Z
4	1	10
4	2	10
5	NULL	11
6	NULL	12

Äußerer Verbund in SQL (4)

- Bedingungen für die linke Tabelle machen in einem linken äußeren Verbund wenig Sinn.
- Z.B. betrachte man diese Anfrage:

```
SELECT A.ATYP, A.ANR, B.SID, B.PUNKTE
FROM   AUFGABEN A LEFT OUTER JOIN BEWERTUNGEN B
       ON A.ATYP = 'H' AND B.ATYP = 'H'
          AND A.ANR = B.ANR
```

- Aufgabe:
  Wird `A.ATYP = 'Z'` in der Ausgabe auftauchen?

  ☐ ja ☐ nein

Inhalt

Präsenzaufgabe: UNION oder Outer Join

- Verwenden Sie das Schema „empdept_public“ im Adminer:
  - dept(deptno, dname, loc)
  - emp(empno, ename, job, mgr°→emp, hiredate, sal, comm°, deptno°→dept)
- Geben Sie Nummer und Namen aller Abteilungen aus, zusammen mit der Anzahl der Beschäftigten in dieser Abteilung. Dabei sollen auch Abteilungen mit 0 Angestellten aufgelistet werden.

deptno	dname	count
10	ACCOUNTING	3
20	RESEARCH	4
30	SALES	6
40	OPERATIONS	0

Bitte keine Lösungen mit Unteranfragen unter SELECT!