1 Deanonymisierung (5 Punkte)

Betrachten Sie die folgenden beiden Tabellen. Tabelle 1 zeigt fiktive anonymisierte Daten eines sozialen Netzwerkes, Tabelle 2 soll einer Webseite für Film- und Serienrezensionen entnommen sein. Gehen Sie davon aus, dass alle Personen aus Tabelle 2 auch in Tabelle 1 vorkommen.

Name	Geschlecht	Alter	Stadt	Lieblingsfilm	Lieblingsserie	Beziehungsstatus
*	weiblich	20-23	Neunkirchen	Scream	Rick und Morty	keine Angabe
*	männlich	24-27	Zweibrücken	Star Wars 8	Game of Thrones	in einer Beziehung
*	weiblich	20-23	Marpingen	El Camino	The Big Bang Theory	in einer Beziehung
*	weiblich	24-27	Saarbrücken	El Camino	Moon Knight	keine Angabe
*	männlich	24-27	Zweibrücken	Die Verurteilten	Game of Thrones	Single
*	männlich	20-23	Merzig	Spider-Man 2	Breaking Bad	in einer Beziehung

Tabelle 1: Anonymisierte Daten eines sozialen Netzwerkes.

Name	Film/Serie	Rezensionsdatum	Sternebewertung
David	Game of Thrones	17.08.2020	★★★★★
Dan	Scream	27.11.2019	★★★★
Lisa	Scream	20.03.2015	★★★★★
Camilla	El Camino	05.04.2022	★★★★
Dan	Game of Thrones	01.03.2019	★★★★★
Camilla	Spider-Man 2	03.09.2019	★★★★★
Lisa	Breaking Bad	03.02.2016	★★★★★
Tobias	Spider-Man 2	10.05.2018	★★★★
Tina	The Big Bang Theory	15.03.2017	★
David	Breaking Bad	27.06.2018	★★
Lisa	El Camino	06.07.2020	★
Dan	Spider-Man 2	01.07.2018	★

Tabelle 2: Daten einer Website für Film- und Serienrezensionen.

(a) Welche persönlichen Informationen erhalten Sie über die folgenden Personen durch Verknüpfen der beiden Datenquellen? Erklären Sie Ihr Vorgehen.

1. Tobias
2. David

(b) Ordnen Sie den weiblichen Personen aus Tabelle 1 Namen zu. Begründen Sie Ihr Vorgehen.

Lösung:

(a) 1. 1 Punkt (0,5 Punkte für den korrekten Tabelleneintrag + 0,5 Punkte für eine korrekte Begründung)

Wir können Tobias vollständig deanonymisieren. Dies ist möglich, da einerseits lediglich zwei Personen Spider-Man 2 eine sehr gute Bewertung gegeben haben - Camilla und Tobias. Da Spider-Man 2 nur eine einzige Person als Lieblingsfilm angegeben hat, ist dies ein starker Hinweis auf Tobias. Die einzige männliche Person, die eine Review zu Breaking Bad erstellt hat, ist David, wobei er lediglich 2 Sterne vergeben hat, dies also vermutlich nicht seine Lieblingsserie sein kann. Da Dan zusätzlich eine negative Bewertung für Spider-Man 2 abgegeben hat, ist Tobias eindeutig der sechsten Reihe zuzuordnen.

2. 2 Punkte (1 Punkt für die korrekten Informationen + 1 Punkt für eine korrekte Begründung)

Sowohl David und Dan haben eine sehr gute Review für Game of Thrones veröffentlicht, sodass beide bzgl. ihrer Lieblingsserie nicht zu unterscheiden sind. Da weder zu Die Verurteilen noch zu Star Wars 8 eine Review existiert, ist es nicht möglich, beide in der Tabelle eindeutig zu identifizieren. Die einzigen Informationen, die für die beiden verbleibenden männlichen Tabelleneinträge gleich sind, ist das Alter von 24-27 Jahren, sowie die Herkunft aus Zweibrücken.

(b) 2 Punkte (1 Punkt für die korrekten Zuordnungen + 1 Punkt für eine korrekte Begründung)

Lisa hat als einzige Frau den Film Scream mit 5 Sternen bewertet, daher gehört der erste Tabelleneintrag sehr vermutlich zu ihr, da sie zudem El Camino, der Lieblingsfilm der anderen beiden Frauen, mit einem Stern bewertet. Folglich können sowohl Camilla, als auch Tina, lediglich über ihre Serie unterschieden werden. Da Tina The Big Bang Theory, die Lieblingsserie des 3. Eintrags, schlecht findet, ist Camilla dem dritten Eintrag und Tina dem vierten Eintrag der ersten Tabelle zuzuordnen.

2 Krankenkassen und Datenzugriff (5 Punkte)

In dieser Aufgabe beschäftigen wir uns mit einer Krankenkasse und ihren personenbezogenen Daten. Normalerweise speichert sich die Krankenkasse lediglich einige Informationen zu ihren Kunden und Kundinnen, basierend auf folgendem Relationenschema:

[Kunden/Kundinnen] : {[KID: int, Geburtsjahr: int, Monatsbeitrag: float, Beitrittsdatum: date]}
[Behandlungen] : {[BID: int, Beschreibung: text]}
[erhalten] : {[Patient*in:(Kunden/Kundinnen→KID), Behandlung: (Behandlungen→BID), Datum: date, Kosten: float]}

Durch einen kürzlichen Datenleak hat die Krankenkasse nun allerdings auch (illegalerweise) Zugriff auf externe Datenbanken bekommen, die das Kaufverhalten ihrer Kunden und Kundinnen beschreiben.

[Artikel] : {[AID: int, Name: string, Preis: real, Nährwerte: real]}
[kaufen] : {[Kunde/Kundin:(Kunden/Kundinnen→KID), Artikel:(Artikel→AID), Datum: date, Menge: float]}

1. Wie kann die Krankenkasse diese Daten, die sie unrechtmäßig erhalten hat, im Zusammenhang mit den bereits vorhandenen Daten missbrauchen, um die Beiträge ihrer Kunden und Kundinnen gewinnbringend anzupassen? (3 Punkte)

Lösung: 3 Punkte
Zunächst kann man die Kunden und Kundinnen berechnen, die die Krankenkasse mehr Geld kosten als sie ihr einbringen. Dies kann allein über die Daten der Krankenkasse realisiert werden, indem man die Kosten aller Behandlungen mit den Monatsbeiträgen seit Beitritt verrechnet. Anschließend kann man die verlustbringenden Kunden und Kundinnen auf ungesunde Lebensweisen untersuchen (Zigaretten, Alkohol, übermäßig kalorienreiches Essen) und bei diesen entsprechend die Beiträge erhöhen, da hier ein erhöhtes Risiko für weitere Behandlungen besteht.

2. Formulieren Sie eine oder mehrere umgangssprachliche Anfragen, die potenzielle Kunden und Kundinnen identifizieren, für die die Krankenkasse die Beiträge anpassen sollte. (2 Punkte)

Lösung: 2 Punkte

- Die KID der Kunden und Kundinnen, deren Behandlungskosten größer sind als alle gezahlten Monatsbeiträge seit dem Beitritt.
- Die KID der verlustbringenden Kunden und Kundinnen aus der ersten Anfrage, die ungesunde Lebensmittel wie Zigaretten und Alkohol und übermäßig kalorienreiches Essen kaufen.

Bei der zweiten Anfrage werden keine konkreten Lebensmittel vorausgesetzt. Darüberhinaus werden alle sinnvollen Anfragen/Ideen akzeptiert.

3 Von SQL zum logischen Plan (5 Punkte)

In dieser Aufgabe möchten wir eine SQL Anfrage, die auf dem nachfolgenden Schema basiert, mit den in der Vorlesung vorgestellten Regeln optimieren.

[Personen] : {[PID: int, Name: string, Geburtsjahr: int, Wohnort: string]}
[Schüler*innen] : {[SID:(Personen→PID), Klassenstufe: int, Klassenraum: int, Schulform: string]}
[Klausuren] : {[KID: int, Fach: string, Thema: string, Dauer: int]}
[Lehrer*innen] : {[LID:(Personen→PID), Hauptfach:string, Gehalt:int, Dienstjahre:int]}
[unterrichten] : {[Schüler*in:(Schüler*innen→SID), Datum:date, Uhrzeit:time,
Lehrer*in:(Lehrer*innen→LID), Fach: string)]}
[korrigieren] : {[Klausur:(Klausuren→KID), Schüler*in:(Schüler*innen→SID),
Lehrer*in:(Lehrer*innen→LID), Note: int]}

```
SELECT Name, Wohnort, Hauptfach, Gehalt
FROM   Personen, Lehrer*innen, unterrichten
WHERE  PID = LID
       AND Lehrer*in = LID
       AND Fach = 'Deutsch'
       AND Geburtsjahr < 2000
       AND Gehalt >= 500;
```

(a) Übersetzen Sie die SQL Anfrage kanonisch in einen Ausdruck der relationalen Algebra. Verwenden Sie dabei nur Projektionen, Selektionen und kartesische Produkte.

(b) Zeichnen Sie den logischen Plan der Anfrage als Baum.

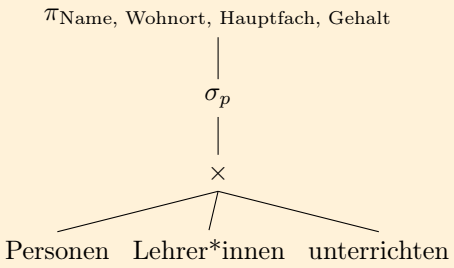
(c) Wenden Sie die aus der Vorlesung und dem Notebook Rule-based Optimization.ipynb bekannten Regeln zur heuristischen Anfrageoptimierung an und zeichnen Sie den optimierten, logischen Anfragebaum, der sich dadurch ergibt.

Lösung:

(a) Vorgesehen ist 1 Punkt (je 0,5 Punkte Abzug für inkorrekte (Join-)Prädikate, inkorrekte Bezeichner durch Verwechslung von Umbenennung und Teilergebnisnotation, unnötige Operatoren durch nicht-kanonische Übersetzung, usw.). Die Reihenfolge, in welcher das kartesische Produkt angewendet wird, spielt hier keine Rolle.

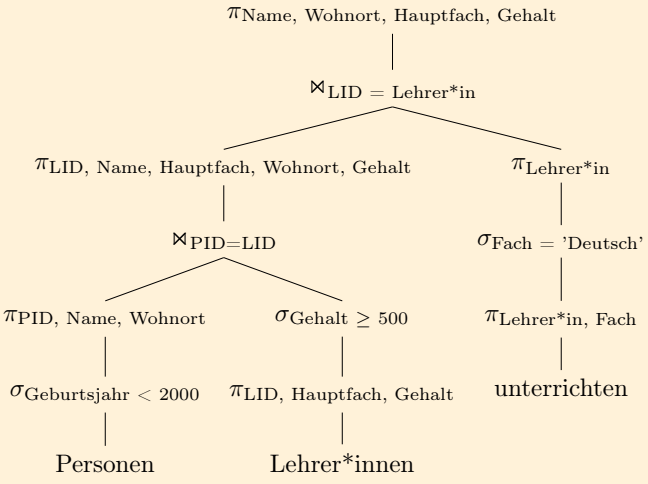
$\pi_{\text{Name, Wohnort, Hauptfach, Gehalt}}\ (\sigma_p\ (\text{Personen} \times \text{Lehrer*innen} \times \text{unterrichten}))$
wobei
$p :=$ PID=LID $\wedge$ Lehrer*in=LID $\wedge$ Gehalt$\geq$500 $\wedge$ Fach='Deutsch' $\wedge$ Geburtsjahr $<$ 2000

(b) Vorgesehen ist 1 Punkt (je 0,5 Punkte Abzug für inkorrekte (Join-)Prädikate, unnötige Operatoren durch nicht-kanonische Übersetzung, usw.). Die Reihenfolge, in welcher das Kreuzprodukt angewendet wird, spielt hier keine Rolle.

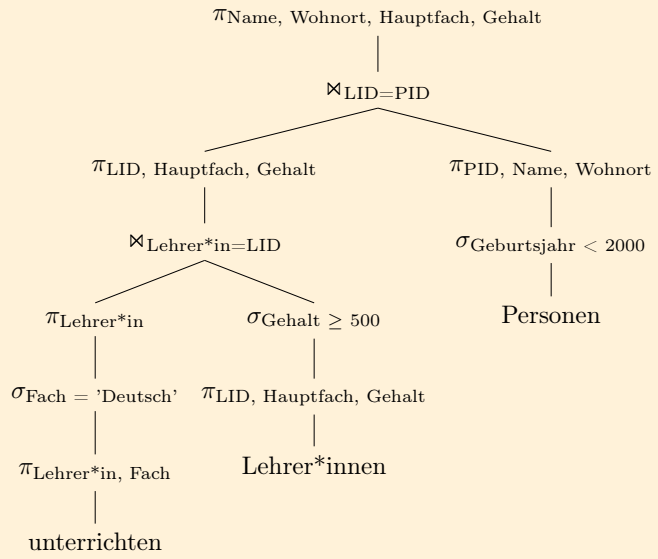


wobei
$p :=$ PID=LID $\wedge$ Lehrer*in=LID $\wedge$ Gehalt$\geq$500 $\wedge$ Fach='Deutsch' $\wedge$ Geburtsjahr $<$ 2000

(c) Vorgesehen sind 3 Punkte (je 0,5 Punkte Abzug für kartesische Produkte, fehlender Predicate Pushdown d.h. auch ein dreistelliger Join widerspricht diesem, überflüssige Attribute durch fehlende Projektionen, usw.).



oder


$\pi_{\text{Name, Wohnort, Hauptfach, Gehalt}}$
$\bowtie_{\text{LID=PID}}$
$\pi_{\text{LID, Hauptfach, Gehalt}}$
$\pi_{\text{PID, Name, Wohnort}}$
$\bowtie_{\text{Lehrer*in=LID}}$
$\sigma_{\text{Geburtsjahr} < 2000}$
$\pi_{\text{Lehrer*in}}$
$\sigma_{\text{Gehalt} \geq 500}$
Personen
$\sigma_{\text{Fach = 'Deutsch'}}$
$\pi_{\text{LID, Hauptfach, Gehalt}}$
$\pi_{\text{Lehrer*in, Fach}}$
Lehrer*innen
unterrichten

4 Kommissar Equi-Join's schwerster Fall (5 Punkte)

In dieser Aufgabe schlüpfen Sie in die Rolle von Juniorkommissar Gruppierung. Nutzen Sie zum Lösen der Aufgabe das beigefügte Notebook. Der mitgelieferte Datensatz ist eine Erweiterung der Daten des NSA.ipynb Notebooks.

[households] : {[id: int, street: string, postcode: int, city: string, floor: int]}
[citizens] : {[id: int, firstname: string, lastname: string, birthday: string]}
[livingIn] : {[citizen_id:(citizens→id), start: string, until: string, household_id:(households→id)]}
[articles] : {[id: int, label: string, unit: string]}
[nutritionalValues] : {[id:(articles→id), calories: int]}
[purchases] : {[article_id:(articles→id), citizen_id:(citizens→id), date: string, amount: real]}

Kommissar Equi-Join sprach neulich mit seinem Vorgesetzten und Mentor, Oberkommissar Theta-Join, über alte Fälle, wodurch Equi-Joins Aufmerksamkeit auf einen seiner wenigen, ungelösten Fälle gelenkt wurde. Hierbei geht es um einen Mord an der Person John Doe, welcher sich am 24.11.1943 ereignete. Gemäß dem Autopsiebericht wurde als Todesursache ein sehr seltenes (und fiktives!) Gift festgestellt, das allerdings laut der zuständigen Rechtsmedizinerin, Dr. Selektion, durch eine Liste an Alltagslebensmittel hergestellt werden kann. Diese Liste besteht aus:

- Genau 500 Gramm Gewürzgurken.
- Mindestens zwei Kilogramm Salat.
- Mindestens ein Kilogramm Karotten, aber maximal (inklusive) drei Kilogramm.

Dr. Selektion meinte noch, dass diese Lebensmittel maximal 5 Tage (inklusive) nach dem Einkauf als Gift verwendet werden können, da ansonsten die Wirkung zu schwach wäre.
Basierend auf dieser Annahme, leitete Kommissar Equi-Join damals eine Überprüfung der dokumentierten Einkäufe in den örtlichen Supermärkten ein. Leider konnte er mit diesen Daten nicht viel anfangen. Daher befragte er anschließend Zeugen und Zeuginnen nach verdächtigen Aktivitäten am Tag des Mordes, jedoch blieb auch dies erfolglos.

Kommissar Equi-Join ist nun allerdings davon überzeugt, mit Ihrer Hilfe den Fall lösen zu können. Dazu holt er die alten Befragungen der Zeugen und Zeuginnen heraus. Leider sind die Aussagen aufgrund des Alters der Dokumente zum größten Teil unlesbar geworden. So kann er lediglich eine Seite mit Informationen finden, in denen Zeugen und Zeuginnen davon berichten, wie verdächtige Personen am Tag des Mordfalls in ihre Haushalte zurückkehren. Dabei sind die folgenden Informationen über die Adressen der Verdächtigen noch lesbar:

- Adresse 1: ...18
- Adresse 2: ...straße 1...
- Adresse 3: L...

Kommissar Equi-Join gibt Ihnen in diesem Zusammenhang das damalige Einwohnerregister, in dem Informationen über die örtlichen Bewohner und deren gemeldeten Häuser zu finden sind, sowie die Daten über die damals registrierten Einkäufe. Können sie aufgrund dieser Daten Kommissar Equi-Join helfen, seinen alten Fall zu lösen? Geben Sie ihre Lösung als SQL-Anfrage ab, die die folgende Ausgabe hat:

- Die Vornamen der Verdächtigen als 'Vorname'.
- Die Nachnamen der Verdächtigen als 'Nachname'

Sie dürfen zum Lösen dieser Aufgabe Unteranfragen sowie Views benutzen. Erläutern Sie zudem in der `jupyter.txt`, ob Sie anhand der ausgegebenen Daten eine*n Hauptverdächtige*n eindeutig identifizieren können.

Lösung: 5 Punkte
Die Idee hinter den Anfragen ist, dass man zunächst allgemein eine Query (View) erstellt, in der man auflistet, wer die genannten Lebensmittel in der Mindestmenge gekauft hat.
Anschließend kann man in einer Anschlussquery die Personen herausfiltern, die alle diese Lebensmittel gekauft haben und diese Personen mit denen vergleichen, die in den beschriebenen Adressen wohnen.
Am Ende sollte eine einzige Person herausgefiltert werden, die den Namen 'Norman Bates' trägt, welche zum Hauptverdächtigen wird.

Abgabe

Lösungen sind in Teams von 2 bis 3 Studierenden bis zum 2. Juni 2022, 10:15 Uhr über Ihre persönlichen Statusseite im CMS einzureichen. Nutzen Sie hierfür die Team Groupings Funktionalität im CMS.

Ihre Abgabe muss dem folgenden Format entsprechen:

```
abgabe.zip
├── abgabe.pdf
└── jupyter.txt
```

Hierbei enthält `abgabe.pdf` Ihre Lösungen zu Aufgabe 1, 2 und 3 und `jupyter.txt` Ihre Lösung zu Aufgabe 4. Achten Sie darauf, dass Sie nur die von Ihnen zu ergänzenden Jupyter Zellen so kopieren, dass Einrückung und Formatierung korrekt sind.

Abgaben, die nicht den oben angegeben Vorgaben entsprechen, führen zu Punktabzug. Einzelabgaben werden nicht mehr korrigiert und mit 0 Punkten bewertet.