Informationssysteme und Datenanalyse
SS 2015
Klausur 21.09.2015

Angaben

Nachname: _____

Vorname: _____

Matrikelnummer: _____

Fakultät / Studiengang (bitte ankreuzen):

- O Fak IV - Bachelor Informatik
- O Fak IV - Bachelor Wirtschaftsinformatik
- O Fak IV - Bachelor Technische Informatik
- O Fak IV - StuPO 90 Informatik
- O BSc – Wirtschaftsingenieurwesen
- O BSc – Wirtschaftsmathematik
- O BSc – Mathematik
- O MSc – Wirtschaftsingenieurwesen
- O MSc – Wirtschaftsmathematik
- O MSc – Mathematik
- O Andere_____

Auf jedem Blatt bitte Namen und Matrikelnummer angeben!

Organisatorisches

Bearbeitungszeit: 75 Minuten
Erreichbare Punkte: 70
Zugelassene Hilfsmittel: Nur ein Wörterbuch (kein elektronisches!)
Weitere Hilfsmittel sind nicht zugelassen.

Aufgabe	Punkte	Erreicht	Korrektor
1	9,5		
2	5		
3	9		
4	6		
5	12		
6	3		
7	8		
8	17,5		
Punktsumme	70		

Aufgabe 1: Relationales Datenmodell/Reverse Engineering (9,5 Punkte)

Gegeben sei ein Informationsmodell zur Planung und Organisation von Veranstaltungen für Unternehmen. Folgende Tabellen sind in einer relationalen Datenbank umgesetzt worden, von der man hier einen konkreten Zustand sieht.

Mitarbeiter	MID	Name	Vorgesetzter
	3	Lundquist	5
	4	Johansson	5
	5	Carlsson	NULL
	6	Lindström	7
	7	Müller	NULL

Kunde	KID	Name
	5	Mia
	6	Emma
	7	Leon
	8	Felix
	9	Helga

WirktMit	MID	VID
	4	1
	5	1
	6	2
	7	2
	4	3

liefert	VID	LID
	3	L3
	3	L1
	1	L1
	2	L2

Veranstaltung	VID	VerantwortlicherMitarbeiter	Bezeichnung	KID
	1	4	Tagung	5
	2	6	Seminar	7
	3	4	Symposium	7

Lieferant	LID	Ort	Entfernung	Lieferung			
				ArtikelNr	Bezeichnung	Anzahl	Bearbeiter
	L1	Hamburg	230	A1	Rotwein	40	Meier
				A2	Bier	20	Schmidt
				A3	Wasser	60	Ruth
	L2	Berlin	30	A1	Rotwein	30	Meier
				A4	Wasser	25	Ursula
				A5	Cola	30	Horst
	L3	München	600	A6	Cola	25	Erika

Aufgabe 1.1 (7,5 Punkte)

Rekonstruieren Sie aus den vorliegenden Tabellen das zugrundeliegende ER-Typ-Diagramm (kein Glossar erstellen!). Versuchen Sie aus den Tabellen möglichst gut auf die Integritätsbedingungen im Original-Modell zurückzuschließen.

Rekonstruiertes ER-Typ-Diagramm:

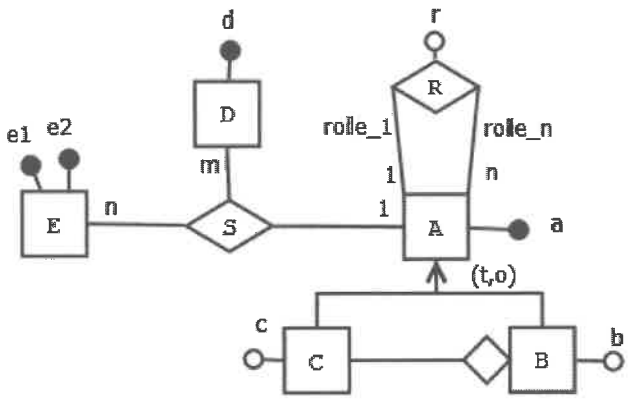
Aufgabe 1.2 (2 Punkte)

Erweitern Sie das in 1.1 aufgestellte EER-Modell um die folgenden Anforderungen und ergänzen Sie dabei alle notwendigen Modellierungskonstrukte im Diagramm:

a. Das Unternehmen beschäftigt *unter anderem* Techniker und Servicemitarbeiter. Bei den Technikern wird zusätzlich die Telefonnummer gespeichert. Bei Servicemitarbeitern wird zusätzlich angegeben, ob es sich um eine Hilfskraft oder einen Festangestellten handelt.
b. Veranstaltungen finden immer zu einer Anfangs- und einer Endzeit in einem Raum statt. Bei den Räumen wird der Preis pro Stunde und die Anzahl der Plätze vermerkt.

Aufgabe 2: Relationaler Datenbankentwurf (5 Punkte)

Gegeben sei das folgende abstrakte EER-Diagramm.

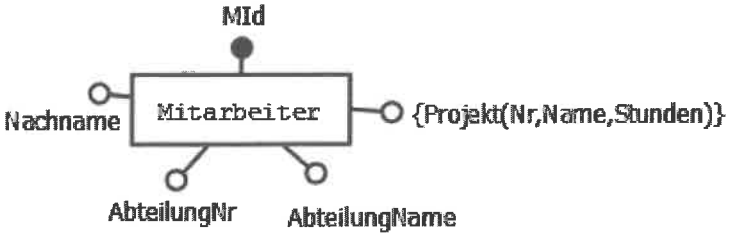


Aufgabe:

Bilden Sie das gegebene abstrakte EER-Diagramm in ein relationales Schema ab. Prümärschlüssel müssen als solche gekennzeichnet werden, Fremdschlüssel müssen vorhanden sein, müssen aber nicht als solche gekennzeichnet werden. Fassen Sie die Relationen soweit wie möglich zusammen.

Aufgabe 3: Normalisierung (9 Punkte)

Gegeben sei folgendes ER-Diagramm mit einem Entity-Typ und eine Bespielinstanz der Relation:



Mitarbeiter	MId	Nachname	AbteilungNr	AbteilungName	Projekt		
					Nr	Name	Stunden
	1	Winter	8	Test	3	P3	15
					6	P6	20
					9	P9	25
	2	Sommer	9	Entwicklung	3	P3	20
					8	P8	10
					9	P9	5
	3	Frühling	8	Test	8	P8	30

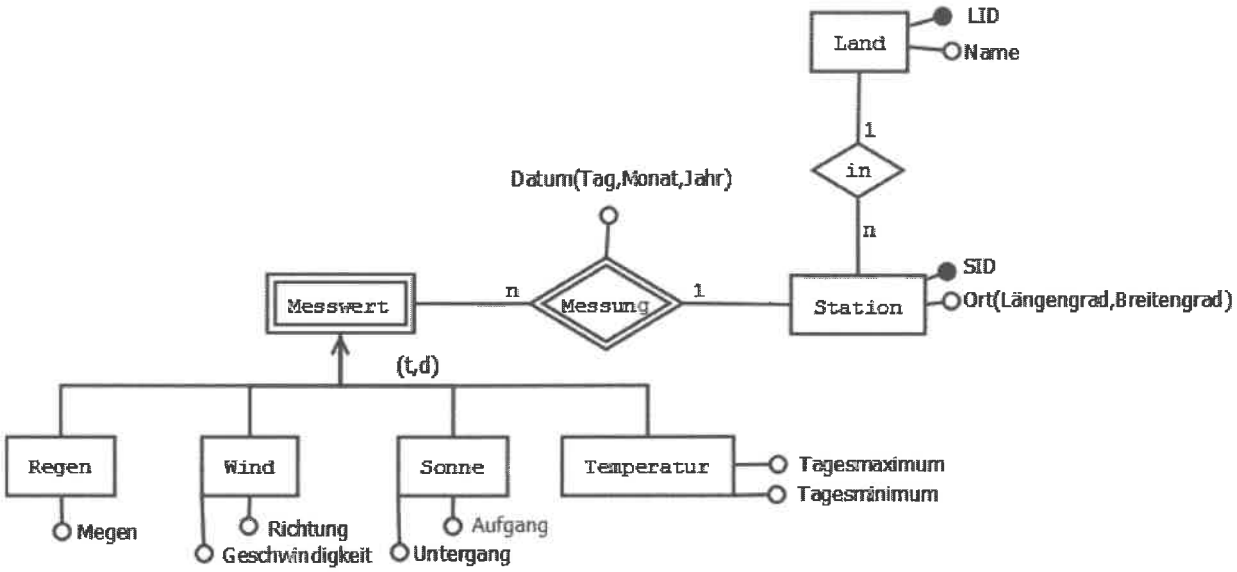
Aufgabe

Die Relation '*Mitarbeiter*' ist nicht in der 1. NF. Normalisieren Sie die Relation bis zur BCNF. Sie müssen bei der Zerlegung die Instanzen nicht angeben, es reicht also das Relationen-Schema. Primärschlüssel müssen gekennzeichnet werden. Fremdschlüssel müssen ebenso vorhanden sein, müssen aber nicht als solche gekennzeichnet werden. Es müssen auch in jedem Schritt die funktionalen Abhängigkeiten angegeben werden, die die jeweilige Normalform verletzen.
Hinweis: 'Stunden' steht für die Anzahl Stunden des jeweiligen Mitarbeiters für das jeweilige Projekt!

Aufgabe 4: Relationale Algebra (6 Punkte)

Gegeben sind das folgende EER-Typ-Diagramm und ein dazugehöriges relationales Schema:

ER-Diagramm



Relationales Schema:

Land(LID, Name)
Station(SID, Längengrad, Breitengrad, LID →Land)
Messwert(SID → Station, Tag, Monat, Jahr, Typ, Regenmenge, Windrichtung, Windgeschwindigkeit, Sonnenaufgang, Sonnenuntergang, Tagesmaximum, Tagesminimum)
$Typ \in \{ Regen, Wind, Sonne, Temperatur \}$

Zur Vereinfachung können folgende Abkürzungen für die Messwertattribute verwendet werden:

Regenmenge **r**, Windrichtung **wr**, Windgeschwindigkeit **wg**, Sonnenaufgang **sa**, Sonnenuntergang **so**, Tagesmaximum **tm**, Tagesminimum **tmi**.

SID	Tag	Monat	Jahr	Typ	R	Wr	Wg	Sa	So	Tm	Tmi
1	21	09	2015	Regen	**5**	*NULL*	*NULL*	*NULL*	*NULL*	*NULL*	*NULL*
2	21	09	2015	Wind	*NULL*	**Nord**	**12**	*NULL*	*NULL*	*NULL*	*NULL*
1	20	09	2015	Wind	*NULL*	**West**	**4**	*NULL*	*NULL*	*NULL*	*NULL*
1	20	09	2015	Sonne	*NULL*	*NULL*	*NULL*	**0800**	**1700**	*NULL*	*NULL*
...											

Instanz der Messwerttabelle.

Aufgabe

1. Wie viele Messungen gibt es aus Frankreich?(1 Punkt)

2. Was war die höchste Windgeschwindigkeit, die je gemessen wurde?(1 Punkt)

3. Wie lang sind Tage in Deutschland durchschnittlich je Monat? (*Sonnenuntergang – Sonnenaufgang)*(1 Punkt)

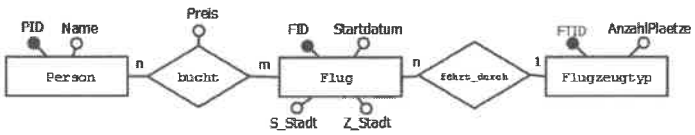
4. Wie viele Stationen haben noch nie Regenmengen gemessen?(1,5 Punkte)

5. Geben sie alle Stationen(SID) an, die in jedem Jahr mindestens einen Messwert haben.(1,5 Punkte)

Aufgabe 5: SQL Anweisungen (12 Punkte)

Gegeben sind das folgende EER-Typ-Diagramm und ein dazugehöriges relationales Schema mit einer Beispielinstanz.

EER-Diagramm



Relationales Schema mit Beispieldatensätze:

Person

PID	Name
1	Till
2	Franka
3	Karoline

bucht

PID	FID	Preis
2	4	300
3	4	280
2	6	100

Flug

FID	Startdatum	FTID	S_Stadt	Z_Stadt
4	03.04.2015	1	Berlin	Madrid
5	06.05.2015	1	Paris	Berlin
6	05.07.2015	2	Hamburg	Rom

Flugzeugtyp

FTID	AnzahlPlaetze
1	130
2	90

Aufgabe

1. Gegeben ist folgende Anfrage in SQL. Geben Sie das Ergebnis dieser Anfrage in tabellarischer Form(Vergessen Sie nicht die Attribute) an: (1 Punkt)

```
SELECT p.PID, AVG(b.PREIS) as X
FROM Person p JOIN Bucht b on p.PID = b.PID
GROUP BY p.PID
HAVING AVG(b.PREIS) < 300;
```

2. Gegeben ist folgende Anfrage in SQL. Geben Sie das Ergebnis dieser Anfrage in tabellarischer Form(Vergessen Sie nicht die Attribute) an und die umgangssprachliche Formulierung dieser Anfrage(Max. zwei Sätze): (3 Punkte)

```
SELECT p.NAME
FROM Person p JOIN Bucht b ON p.PID = b.PID JOIN Flug f ON b.FID = f.FID
WHERE Z_Stadt = 'Madrid' AND
            NOT EXISTS (SELECT *
                        FROM BUCHT B1 JOIN FLUG FL ON B1.FID = FL.FID
                        WHERE Z_STADT !='Madrid' AND b.PID = B1.PID );
```

3. Gegeben ist folgende Anfrage in SQL. Geben Sie das Ergebnis in tabellarischer Form an (Vergessen Sie nicht die Attribute) und die umgangssprachliche Formulierung dieser Anfrage(Max. zwei Sätze): (4 Punkte)

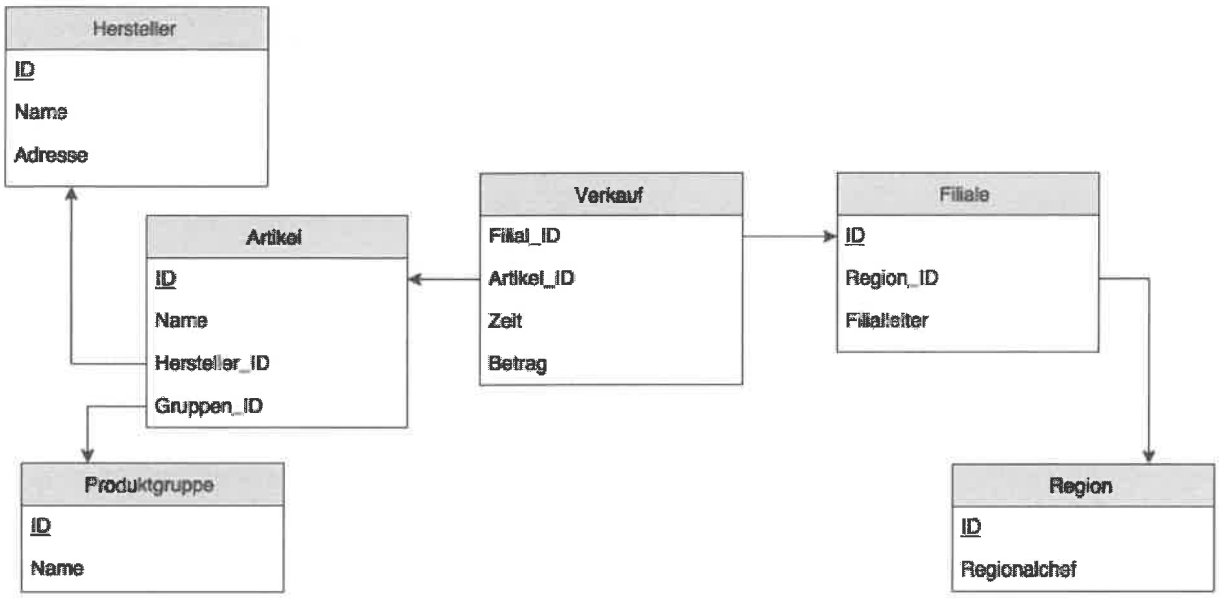
```
SELECT *
FROM (SELECT b.PID
      FROM BUCHT b join FLUG f on b.FID = f.FID
      WHERE f.Z_stadt = 'Madrid') as A,
                                    (SELECT b.PID
                                    FROM BUCHT b join FLUG f on b.FID = f.FID
                                    WHERE f.Z_stadt = 'Rom') as B
WHERE A.PID = B.PID;
```

4. Erstellen Sie die SQL Select-Anweisung, die die folgende Anfrage berechnet: Wie oft ist Franka nach Rom geflogen?.(2 Punkte)

5. Erstellen Sie die SQL Select-Anweisung, die die folgende Anfrage berechnet: Welche Personen(PID, Name) haben keinen Flug gebucht? (2 Punkte)

Aufgabe 6: DWH-Mehrdimensionale Modellierung(3 Punkte)

Betrachten Sie das folgende ER-Schema eines OLAP Würfels:



Aufgabe:

1. Benennen sie die Fakten- und Dimensionstabellen: (0.5 Punkte)

2. In der Vorlesung wurden drei Darstellungen vorgestellt um einen OLAP Würfel auf relationale Tabellen abzubilden. Welcher der vorgestellten Darstellungen entspricht dieses Schema? (0.5 Punkte)

3. Wie sähe das Schema aus falls eine Sternschema-Darstellung verwendet worden wäre? (1 Punkt)

4. Warum wird im Data Warehousing auf spezielle Darstellungen wie das Sternschema gesetzt? Die Daten ließen schließlich auch als normalisiertes ER Schema (z.B. in 3NF oder BCNF) darstellen. (Stichpunkte) (1 Punkt)

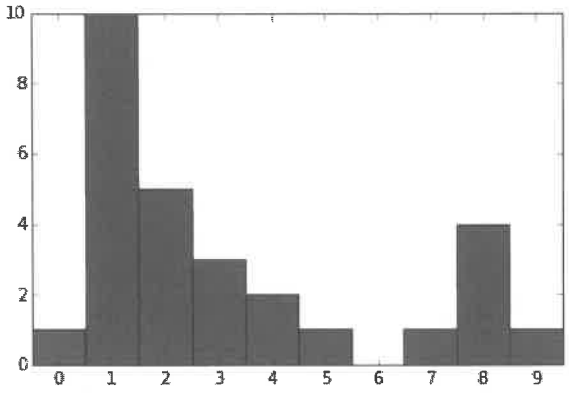
Aufgabe 7: Datenanalyse (8 Punkte)

Aufgabe 7.1 – Begrifflichkeiten: (1 Punkt)

Erläutern Sie (kurz!) den Unterschied zwischen überwachten (supervised) und unüberwachten (unsupervised) Lernverfahren? Geben Sie für beide Verfahren jeweils ein Beispiel an.

Aufgabe 7.2 – Visualisierung und Basisstatistiken: (1.5 Punkte)

Gegeben ist das folgende Histogramm:



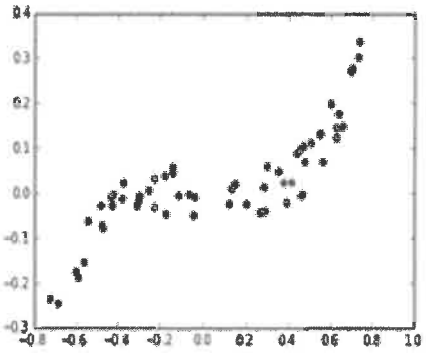
Bestimmen Sie anhand dieses Histogramms folgende Basisstatistiken der Verteilung:

1. Median:

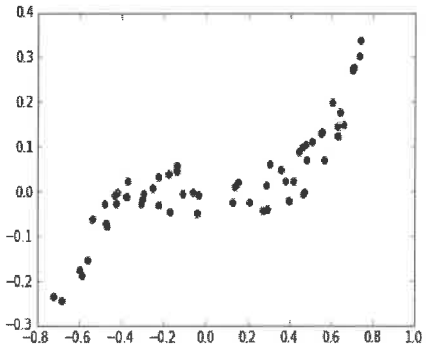
2. Durchschnitt:

3. Modalwert (Modus):

Aufgabe 7.3 – Regression: (1.5 Punkte)



(a)



(b)

a) Skizzieren sie in Bild (a) eine Regressionsfunktion die den Datensatz *underfittet*, in Bild (b) eine Regressionsfunktion die den Datensatz *overfittet* (1P):

b) Nennen Sie eine Möglichkeit um Overfitting beim Trainieren eines Modelles zu vermeiden: (0.5P)

Aufgabe 7.4 – Recommender Systems: (2 Punkte)

Gegeben ist der folgende Datensatz welcher angibt welche Filme von welchen Benutzern angesehen wurden:

	Titanic	Jurrasic Park	Star Wars	Dark Knight	Forrest Gump
Benutzer 1	1	-	1	1	-
Benutzer 2	1	1	-	-	1
Benutzer 3	1	1	-	1	-

a) Welchen Film würde ein *Most-Popular* Recommender System auf Basis dieser Daten vorschlagen? Warum? (0.5 Punkte)

b) Berechnen Sie das Kosinus Ähnlichkeitsmaß (Cosine-Similarity) zwischen den Benutzern 1, 2 & 3 (1 Punkt):

c) Welchen Film würde ein *User-based Collaborative Filtering* Recommender System Benutzer 1 auf Grund dieser Daten vorschlagen? Warum? (0.5 Punkte)

Aufgabe 7.5 – Clustering: (2 Punkte)

Gegeben sind die folgenden sechs Datenpunkte:

$$x_1 = -2, \quad x_2 = 0, \quad x_3 = 2, \quad x_4 = 6, \quad x_5 = 8, \quad x_6 = 10$$

Für diese Daten sollen mittels des K-Means Algorithmus ein Clustering erstellt werden. Nehmen Sie hierzu k=2, und die folgenden Clusterzentren an:

$$c_1 = 3, \quad c_2 = 12,$$

a) Welche Clusterzentren erhalten sie nach der ersten Iteration von K-Means? (1 Punkt)

b) Zu welchen Clusterzentren wird der Algorithmus vermutlich konvergieren? (0.5 Punkte)

c) Sie sollen nun auf einem weiteren Datensatz mittels K-Means ein Clustering bestimmen. Leider ist Ihnen die Struktur der Daten, und daher der optimal Wert von k unbekannt. Um den optimalen Wert zu bestimmen lassen sie K-Means mit verschiedenen Werten für k laufen, und messen die Gesamtkosten des gefunden Clusterings (d.h. den aufsummierten Abstand aller Punkte zu ihren zugewiesenen Clusterzentren). Sie erhalten die folgende Tabelle:

k	1	2	3	4	5	6	7
Kosten	3314	1764	954	229	204	183	163

Welchen Wert für k würden Sie anhand dieser Messungen wählen? Warum? (Hinweis: Es ist hilfreich die Messungen zu skizzieren). (0.5 Punkte)

Aufgabe 8: MC (17,5 Punkte)

Diese Aufgabe umfasst **21 MC-Fragen**. Sie bestehen aus jeweils drei Antwortalternativen, wobei mindestens eine Alternative richtig und mindestens eine Alternative falsch ist. Im Allgemeinen gibt es für jede Aufgabe, in welcher alle korrekten Aussagen markiert worden sind, zwischen 0,5 und 3 Punkte. Wurden nicht alle richtigen Aussagen identifiziert oder falsche Aussagen markiert, wird die jeweilige Aufgabe mit 0 Punkten bewertet.

1. Was bedeutet „links-Totalität"? (0,5 Punkte)
   - ☐ Jede Entity auf der linken Seite muss eine Beziehung haben.
   - ☐ Jede Entity auf der linken Seite kann eine Beziehung haben.
   - ☐ Jede Entity auf der linken Seiten darf maximal eine Beziehung haben.

2. Kardinalitäten in einem ER-Modell bezeichnen,(0,5 Punkte)
   - ☐ wie viele Instanzen von einem Entitytypen existieren können.
   - ☐ wie häufig eine Entity eine Beziehung eingeht.
   - ☐ warum eine Entity eine Beziehung eingeht.

3. Welche Aussagen zu einem Primärschlüssel sind richtig? (0,5 Punkte)
   - ☐ Ein Primärschlüssel muss minimal sein.
   - ☐ Ein Primärschlüssel darf kein Fremdschlüssel sein.
   - ☐ Ein Primärschlüssel-Wert darf nicht NULL sein.

4. Welche der folgenden Mengen von Operatoren besteht nur aus Basisoperatoren? (0,5 Punkte)
   - ☐ $\{\pi, \bowtie, \cap, \sigma, \times, \rho\}$
   - ☐ $\{\sigma, \pi, \cup, \times, -, \rho\}$
   - ☐ $\{/, \times, \pi, \sigma, \bowtie, \cup\}$

5. Welche Probleme können auftreten, wenn n:m-Beziehungen in einer Tabelle aufgeführt werden?(0,5 Punkte)
   - ☐ Es können Redundanzen und Inkonsistenzen entstehen.
   - ☐ Es können „NULL"-Werte in einer Tabelle entstehen.
   - ☐ Es können Schlüsselattribute verlorengehen.

6. Welche Reihenfolgen der folgenden SQL-Anweisungen sind möglich? (0,5 Punkte)
   - ☐ SELECT ... FROM (SELECT ... FROM ... WHERE ... GROUP BY ... HAVING ...)
   - ☐ SELECT ... GROUP BY ... FROM
   - ☐ SELECT ... FROM ... WHERE ... GROUP BY ... HAVING ...

7. Welche Aussagen gelten? (0,5 Punkte)
   - ☐ SQL ist eine deklarative Abfragesprache.
   - ☐ Die Ergebnisse von SQL Anfragen sind im allgemeinen Multimengen.
   - ☐ SQL steht für „SECURE QUERY LANGUAGE".

8. Durch welchen Ausdruck in relationaler Algebra wird die nachstehende SQL-Anfrage richtig ausgedrückt? (1 Punkt)

```
SELECT  P.Name
FROM    Person p JOIN bucht b ON Flug f
WHERE   p.PID=b.PID AND b.FID = f.FID AND b.Preis > 100;
```

- ☐ $\pi_{Name}(\sigma_{\text{Preis}>100}(\sigma_{\text{bucht.FID=Flug.FID}}(\sigma_{\text{Person.PID=bucht.PID}}(\text{Person} \times \text{bucht} \times \text{Flug}))))$
- ☐ $\pi_{\text{Name}}(\text{Person} \bowtie (\sigma_{\text{Preis}>100}(\text{bucht}) \bowtie \text{Flug})$
- ☐ $\pi_{Name}(\text{Person} \times \sigma_{\text{Preis}>100}(\text{bucht}) \times \text{Flug})$

9. Was sind Nachteile der Normalisierung? (0,5 Punkte)
   - ☐ Redundanzen werden minimiert.
   - ☐ Anomalien werden weitestgehend beseitigt.
   - ☐ Schlechtes Laufzeitverhalten bei JOIN mehrerer Relationen.

10. Welche der folgenden Aussagen gelten?(0,5 Punkte)
    - ☐ Die Funktionale Abhängigkeit definiert einen Constraint für alle möglichen Instanzen einer Relation.
    - ☐ Die Funktionale Abhängigkeit definiert einen Constraint für eine konkrete Instanz einer Relation.
    - ☐ Ein Fremdschlüssel bestimmt meistens alle anderen Attribute funktional.

11. Wie lautet die natürlichsprachliche Beschreibung der nachstehenden SQL-Anfrage? (3 Punkte)

```
SELECT a.Vorname, a.Nachname
FROM   Angestellter a
WHERE  NOT EXISTS (
       SELECT *
       FROM   Projekt p
       WHERE  p.Leiter = 'Tom'
       AND    NOT EXISTS (
              SELECT *
              FROM   ArbeitetIn ai
              WHERE  ai.PID = p.PID
              AND    ai.AID = a.AID));
```

- ☐ Finde alle Angestellten, die an allen Projekten, die von Tom geleitet werden, arbeiten.
- ☐ Finde alle Angestellten, die an allen Projekten arbeiten, die nicht von Tom geleitet werden.
- ☐ Finde alle Angestellten, die an irgendeinem Projekt arbeiten, das von Tom geleitet wird.

12. Wenn dieselbe Lese-Anfrage in einer Transaktion nacheinander unterschiedliche Ergebnisse liefert, dann ist das ein Beispiel für ein ... (0,5 Punkte)
    - ☐ Non-Repeatable Read
    - ☐ Phantom-Problem
    - ☐ Dirty Read

13. Folgender Ausschnitt einer Table Definition ist gegeben: (1 Punkt)

```
CREATE TABLE X(
…,
, ref INTEGER FOREIGN KEY REFERENCES Y(id) ON DELETE CASCADE
, …)
```

Die Tabelle Y wird nun komplett gelöscht. Dann gilt:

- ☐ Die Tabelle X wird ebenfalls komplett gelöscht
- ☐ Elemente aus X, bei denen 'ref' gesetzt ist werden ebenfalls gelöscht
- ☐ Das Attribut 'ref' wird aus dem Schema entfernt

14. Gegeben sei die Relation R (A, B, C, D, E) mit den funktionalen Abhängigkeiten: A → B, B → C und (B, C) → D. Welche der folgenden Möglichkeiten ist ein möglicher Primärschlüssel für R? (2 Punkte)
    - ☐ A
    - ☐ (A, E)
    - ☐ (A, C)

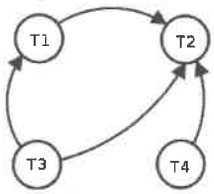
15. Atomizität bedeutet, dass (0,5 Punkte)
    - ☐ Eine Transaktion nur aus einer einzigen Aktion bestehen darf
    - ☐ Eine Transaktion ganz oder gar nicht ausgeführt wird
    - ☐ Eine Transaktion nicht von einer anderen Transaktion unterbrochen werden darf

16. Welche Anomalie könnte in dem gegebenem Schedule auftreten?(0,5 Punkte)

T1	T2
R(A)	
	W(A)
R(A)	

- ☐ Phantom-Problem.
- ☐ Non-Repeatable Write.
- ☐ Non-Repeatable Read.

17. Gegeben sei zu einem Schedule S1 folgender Graph (1 Punkt)



Welche Aussagen sind richtig?

- ☐ S1 ist konfliktserialisierbar, ein konfliktäquivalenter serieller Schedule zu S1 ist: T4 → T3 → T1 → T2.
- ☐ S1 ist konfliktserialisierbar, ein konfliktäquivalenter serieller Schedule zu S1 ist: T3 → T1 → T4 → T2.
- ☐ S1 ist nicht konfliktserialisierbar, da der Graph ein Zyklus enthält.