Raum	
Platz	

Informationssysteme und Datenanalyse

Schriftlicher Test (International Version)

24.09.2018

Dies ist der Test der Lehrveranstaltung *Informationssysteme und Datenanalyse*. Bitte füllen Sie die Tabelle auf diesem Deckblatt aus und unterschreiben Sie den untenstehenden Hinweis.

Hinweise:

- You can answer any question in English or German.
- Die Bearbeitungszeit für diesen Test beträgt 60 Minuten plus 10 Minuten Einlesezeit. Es können in 7 Fragen insgesamt 50 Punkte erreicht werden. Während der Einlesezeit darf **nicht** geschrieben werden.
- Wenn Sie mehr als den zur Bearbeitung einer Aufgabe vorgesehenen Platz benötigen, können Sie ihre Antwort auf einer der freien Seiten fortsetzen. Machen Sie eine Weiterführung ihrer Antwort eindeutig kenntlich.
- Dieser Test besteht aus **15** Seiten. Bitte zählen Sie die Vollständigkeit der Seiten direkt nach Beginn der Einlesezeit.
- Bitte schreiben Sie außerdem direkt nach Beginn der Schreibzeit ihren Namen und ihre Matrikelnummer auf jede Seite.
- Die Verwendung von eigenem Papier ist **nicht** erlaubt. Zusätzliche leere Blätter werden auf Nachfrage ausgeteilt.
- Auf Ihrem Platz dürfen sich lediglich mehrere *dokumentenechte* Stifte sowie ihr Personal- und Studierendenausweis befinden. Einträge mit roten oder grünen Stiften sowie Füller und/oder Bleistift werden nicht gewertet. Weitere Hilfsmittel sind nicht zugelassen. Sämtliche elektronischen Geräte müssen sich ausgeschaltet in Ihrer Tasche befinden. Diese müssen Sie in der Reihe vor Ihnen oder anderweitig entfernt von Ihrem Platz abstellen.
- Klingelnde elektronische Geräte (Smartphones, Smartwatches o.Ä.) gelten als Täuschungsversuch.

Matrikelnummer	
Nachname(n)	
Vorname(n)	
Studiengang	

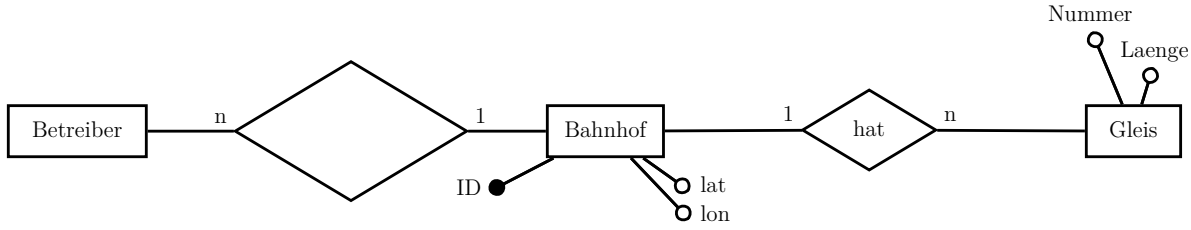
Hiermit bestätige ich, dass ich die oben genannten Hinweise verstanden habe und mich in der Lage fühle, diesen Test durchzuführen.

Unterschrift:

Aufgabe	Punkte	Erreicht	Korrektor
Datenbankentwurf	7		
Relationaler Entwurf	8		
Anfragesprachen	9		
XML	4,5		
Data Streams Management & DWH	5,5		
Data Analysis	6		
Multiple Choice	10		
Summe	50		

Aufgabe 1: Datenbankentwurf (7 Punkte)

Gegeben sei das folgende Entity-Relationship-Diagramm für eine *Bahnhofsdatenbank*. (*Vocabulary: Bahnhof – train station, Gleis – plattform, Betreiber – operator, Haltepunkt – station, Fernbahnhof – main line station, Regiobahnhof – local railway station, Bedarfshalt – flag stop*)



1.1. Vervollständigen Sie das gegebene ER-Diagramm zu einem syntaktisch korrekten (*syntactically correct*) ER-Diagramm. (1,5)

1.2. Ergänzen Sie das gegebene ER-Diagramm um die folgenden Angaben. Achten Sie dabei auch auf mögliche Integritätsbedingungen (*integrity constraints*).

a) Jeder Bahnhof kann mit beliebig vielen Bahnhöfen verbunden sein. Jede solche Verbindung verfügt über eine Liniennummer. (*Every train can be connected to arbitrary many other train stations. A line number is associated with each such connection.*) (1,5)

b) Jedes Gleis muss einem Bahnhof zugeordnet sein. (*Every plattform has to be assigned to a train station.*) (0,5)

1.3. Gegeben seien außerdem die folgenden Relationen. Erweitern Sie das ER-Diagramm aus Aufgabe 1 durch Verwendung eines Abstraktionskonzeptes zu einem erweiterten (*Extended*) ER-Diagramm (EER-Diagramm), das dem gegebenen relationalen Modell entspricht. Nutzen Sie die Informationen aus den gegebenen Relationen. Weitere Tupel als die angegebenen existieren nicht. Achten Sie dabei darauf, dass Ihre Modellierung nicht kapazitätserhöhend (*capacity-increasing*) oder kapazitätsvermindernd (*capacity-decreasing*) ist. (3,5)

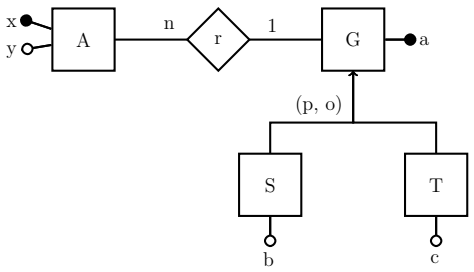
Hinweis: Die Attribute Haltepunkt, Fernbhf und Regiobhf sind binäre Attribute, die den Typ einer Instanz festlegen. (*The attributes Haltepunkt, Fernbhf and Regiobhf are binary attributes indicating the type of an instance.*)

Bahnhof

ID	lat	lon	Haltepunkt	Fernbhf	Regiobhf	Bedarfshalt	v_{pass}	Baecker	{Parkplatz}
XDAR	56.15	10.20	FALSE	TRUE	TRUE	NULL	40	2	{250}
WM	54.49	13.59	FALSE	TRUE	FALSE	NULL	10	0	NULL
HM	52.29	8.93	FALSE	TRUE	TRUE	NULL	110	1	{250, 65}
WKBR	53.15	12.10	TRUE	FALSE	TRUE	TRUE	NULL	NULL	{5}
ADF	53.56	9.99	TRUE	TRUE	TRUE	FALSE	60	3	{}
XOSJV	63.45	10.91	FALSE	FALSE	TRUE	NULL	NULL	NULL	{200, 300, 50}
ZETN	59.44	24.74	FALSE	FALSE	TRUE	NULL	NULL	NULL	{30, 60}
XDLV	55.36	10.59	TRUE	TRUE	FALSE	FALSE	180	0	NULL

Aufgabe 2: Relationaler Entwurf (8 Punkte)

2.1. Gegeben sei das folgende erweiterte Entity-Relationship-Modell.



a) Erstellen Sie nun zum oben gegebenen EER-Diagramm den Relationalen Entwurf (*relational design*). Benutzen Sie zur Abbildung der Generalisierung/Spezialisierung den ER-Stil. (2)

b) Wie viele Relationen entstehen, wenn das gesamte oben gegebene EER-Diagramm im Null-Stil und im objektorienterten Stil abgebildet wird? (1)

Null-Stil		OO-Stil	

c) Die Abbildung der Generalisierung/Spezialisierung im gegebenen ER-Diagramm mit dem ER-Stil ist (0,5)

- ◯ kapazitätserhaltend. (*capacity-preserving*)
- ◯ kapazitätsvermindernd. (*capacity-decreasing*)
- ◯ kapazitätserhöhend. (*capacity-increasing*)

d) Weiterhin ist die Relation A mit $A(\underline{u, w}, x, z)$ sowie den funktionalen Abhängigkeiten (*functional dependencies*) (2,5)

$(u, w) \to x$, $(u, w) \to z$, $z \to u$, $u \to x$

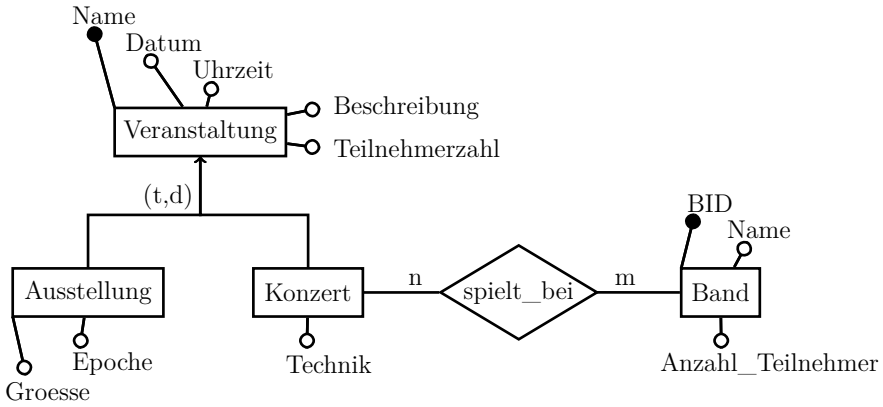
gegeben. Weitere funktionale Abhängigkeiten existieren nicht.

Normalisieren Sie die Relation A bis zur Boyce-Codd-Normalform. Geben Sie für jeden Normalisierungsschritt die Zerlegung (*decompostion*) der Relation(en) sowie die möglicherweise Normalform-verletzende (*violating*) funktionalen Abhängigkeit (*functional dependency*) an. Unterstreichen Sie Schlüsselattribute (*key attributes*). Sie können davon ausgehen, dass sich A bereits in der ersten Normalform befindet.

2.2. Erklären Sie in höchtestens drei Sätzen den Begriff Abhängigkeitstreue (*dependency preservation*). In welche Normalformen kann ein Relationenschema zerlegt werden ohne dass die Abhängigkeitstreue verletzt wird? (2)

Aufgabe 3: Anfragesprachen (9 Punkte)

Gegeben sei folgendes Datenbankschema eines Kulturzentrums, das bereits beispielhafte Tupel enthält.
(*Vocabulary: Veranstaltung - event, Konzert - concert, Ausstellung - exhibition, Teilnehmeranzahl - number of participants*)



Spielt_Bei

BID	VName
1	Open Flair Festival
2	Eurovision Songcontest
2	Open Flair Festival
3	Eurovision Songcontest
4	Musikantenstadl
4	Rammstein Live
5	Rammstein Live
2	Musikantenstadl
5	Open Flair Festival

Band

BID	Bandname	Anzahl_Musiker
1	Rise Against	4
2	Papa Roach	8
3	Radiohead	5
4	Helene Fischer	3
5	Rammstein	6
6	Phil Collins	3

Ausstellung

VName	Epoche	Groesse
Sommerausstellung	Gegenwart	groß
Vernissage Berlin-Mitte	`NULL`	klein

Konzert

VName	Technik
Open Flair Festival	LVX99 Bundle
Musikantenstadl	Soundmaster XL
Eurovision Songcontest	Stereoanlage ZZZ
Rammstein Live	Dosentelefon Nofeletnesod

Veranstaltung

VName	Datum	Uhrzeit	Teilnehmerzahl	Beschreibung
Open Flair Festival	2017-08-09	20:00:00	30000	Rockfestival
Musikantenstadl	2017-03-12	17:00:00	2500	BR-Abendprogramm
Eurovision Songcontest	2016-05-10	20:15:00	9999	Wettbewerb
Rammstein Live	2016-12-07	16:00:00	100	Tourneestart
Vernissage Berlin-Mitte	2017-07-20	08:00:00	42	Hipsterstuff
Sommerausstellung	2017-07-01	08:30:00	1337	Action Painting

3.1. Formulieren Sie eine gültige Anfrage in relationaler Algebra, die das folgende Frage beantwortet: *Welche Bands haben zwischen einschließlich 2 und einschließlich 5 Anzahl_Musiker?* (2)

Relationale Algebra:

SQL:

3.2. Geben Sie das Ergebnis (inklusive Attributnamen) auf Basis der gegebenen Relationen zur folgenden Anfrage an. (2)

```
SELECT SUM(b.Anzahl_Musiker) AS R
FROM Band b
WHERE b.Bandname LIKE '_a%';
```

Ergebnistabelle:

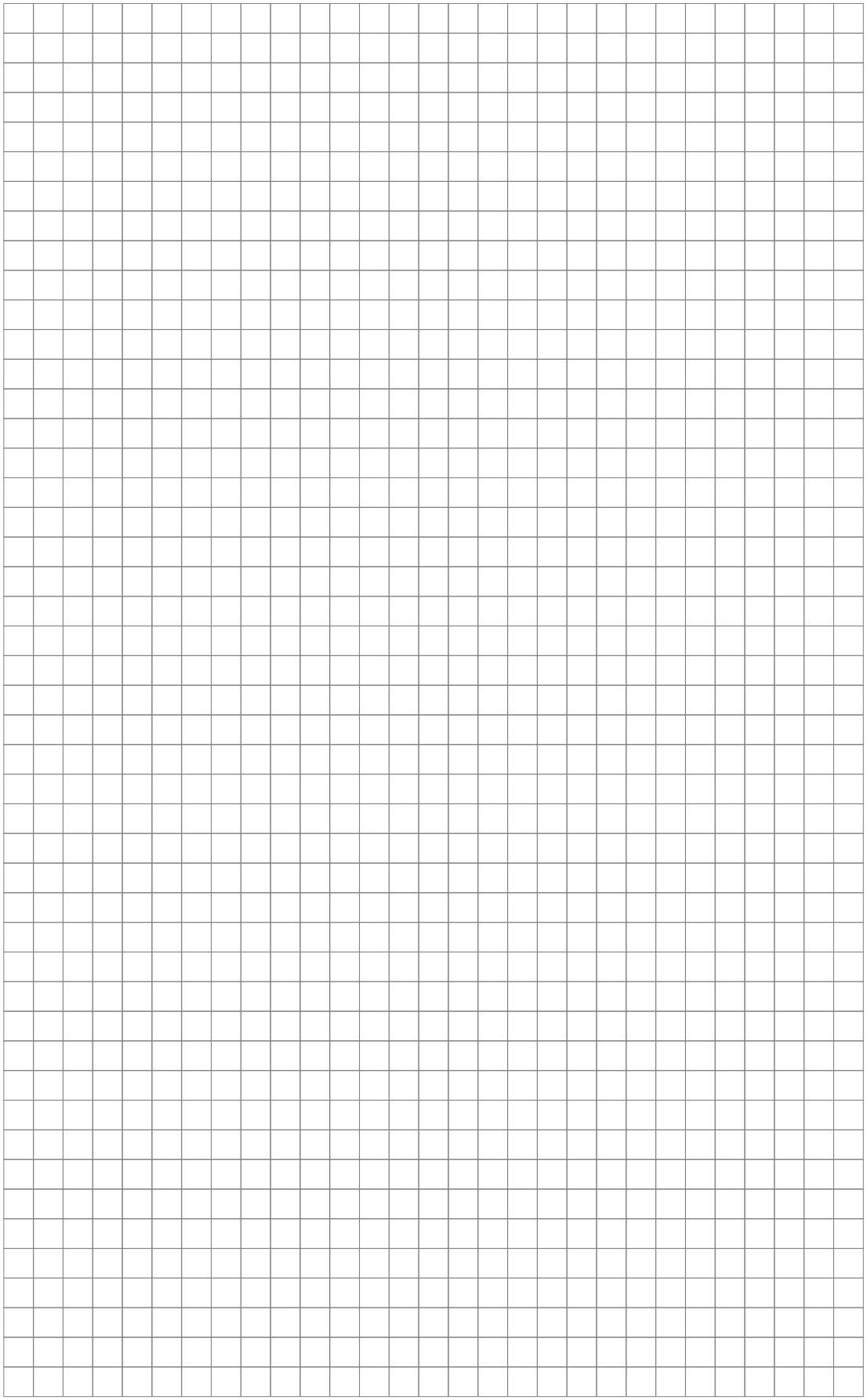
3.3. Formulieren Sie eine gültige Anfrage in relationaler Algebra, die die folgende Frage beantwortet: (3)
Welche Bands haben an mehr als einer Veranstaltung teilgenommen?
Hinweis: Für die Lösungstabelle sind nur die Bandnamen und die Anzahl der Teilnahmen relevant.

Relationale Algebra:

3.4. Geben ist die folgende Anfrage in relationaler Algebra. Schreiben sie die äquivalente SQL-Anfrage dazu. (2)

$\mathbf{Band} \ltimes (\pi_{BID}(\mathbf{Band}) - \pi_{BID}(\mathbf{spielt\_bei}))$

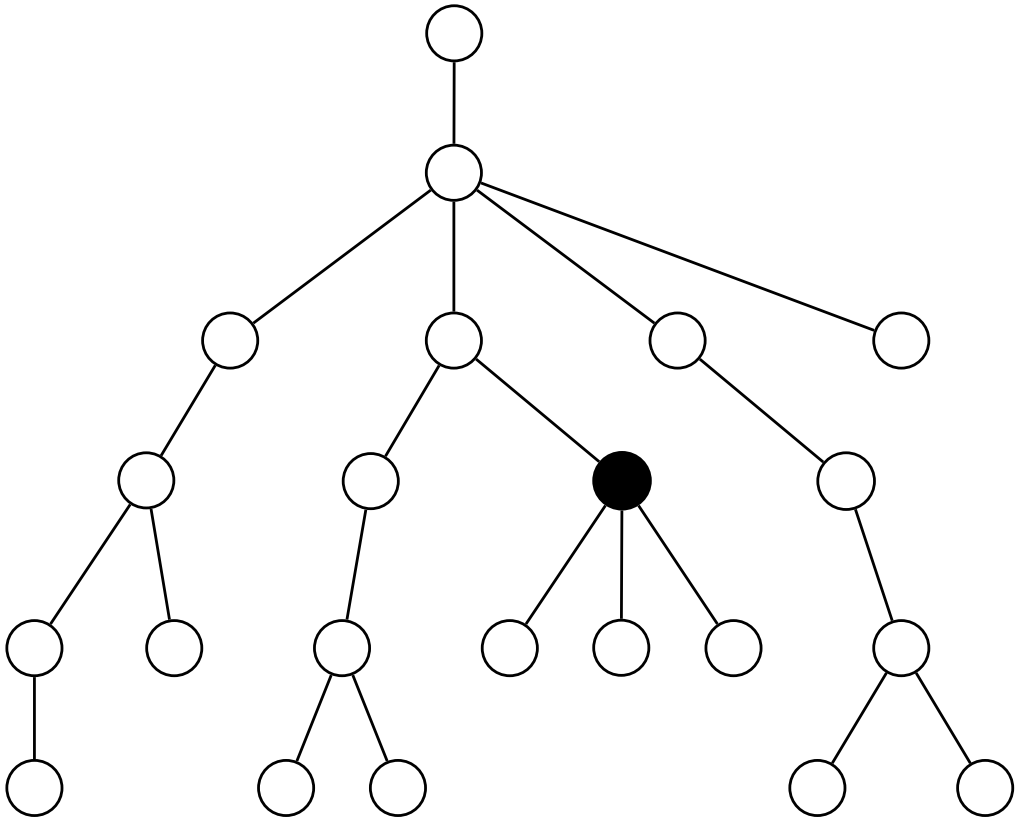
SQL:



Aufgabe 4: XML (4,5 Punkte)

4.1. Gegeben ist die Baum-Repräsentation (*tree representation*) eines XML-Dokuments. Markieren Sie die angegebenen Achsen (*axes*) ausgehend vom schwarz markierten Knoten (*node*), indem Sie die enthaltenen Knoten umkreisen (*(to) circle*) und das Umkreiste mit dem Namen der Achse beschriften. Sollte eine Achse keine Knoten enthalten, brauchen Sie für diese keine Markierung vorzunehmen. (2,5)

`parent`, `descendant-or-self`, `following`, `preceding-sibling`, `self`



4.2. Vervollständigen Sie die unten angegebene allgemeine Syntax für einen Lokalisierungsschritt (*localization step*) in XPath um die fehlenden Bestandteile (*missing components*) in den Lücken (*gaps*). (1)

`_____::_____[_____]`

4.3. Gegeben seien die untenstehenden XPath-Anfragen. Geben Sie wie im Beispiel für jede der Anfragen an, welche Achsen (*axes*) in der Anfrage angesprochen werden. (1)

`/house/flat`	child
`/house/@residents`	
`//item/../sum`	

Aufgabe 5: Data Streams Management & DWH (5,5 Punkte)

5.1. Gegeben sei die folgende Ergebnismatrix eines Count-Min Sketch-Durchlaufs:

$$\begin{array}{c} \\ h_0 \\ h_1 \\ h_2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 17 & 49 & 18 \\ 41 & 29 & 14 \\ 26 & 19 & 39 \end{bmatrix} \end{array}$$

a) Berechnen Sie mit Hilfe des Sketches, wie das folgende Element mit den angegebenen Werten der Hashfunktionen höchstens beobachtet wurde. (1)

v	$h_0(v)$	$h_1(v)$	$h_2(v)$
„Bello“	1	0	2

.....

5.2. Sie verfügen über zwei Bloomfilter, die mit der selben Bitmapgröße (*bitmap size*) b und den gleichen k Hashfunktionen erzeugt wurden. Der erste Bloomfilter wurde über eine Menge (*set*) von Elementen A erzeugt. Der zweite Bloom-Filter wurde über eine Menge von Elementen C erzeugt. Beschreiben Sie in höchstens drei Sätzen, wie die beiden Bloomfilter sinnvoll zu einem einzigen Bloomfilter mit den selben Parametern kombiniert werden können, mit dem Aussagen über Elemente in der Menge $A \cup C$ getroffen werden können. (2,5)

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

5.3. Beschreiben Sie in höchstestens zwei Sätzen, wie ein Snowflake-Schema in ein Fullfact-Schema überführt werden kann. (2)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

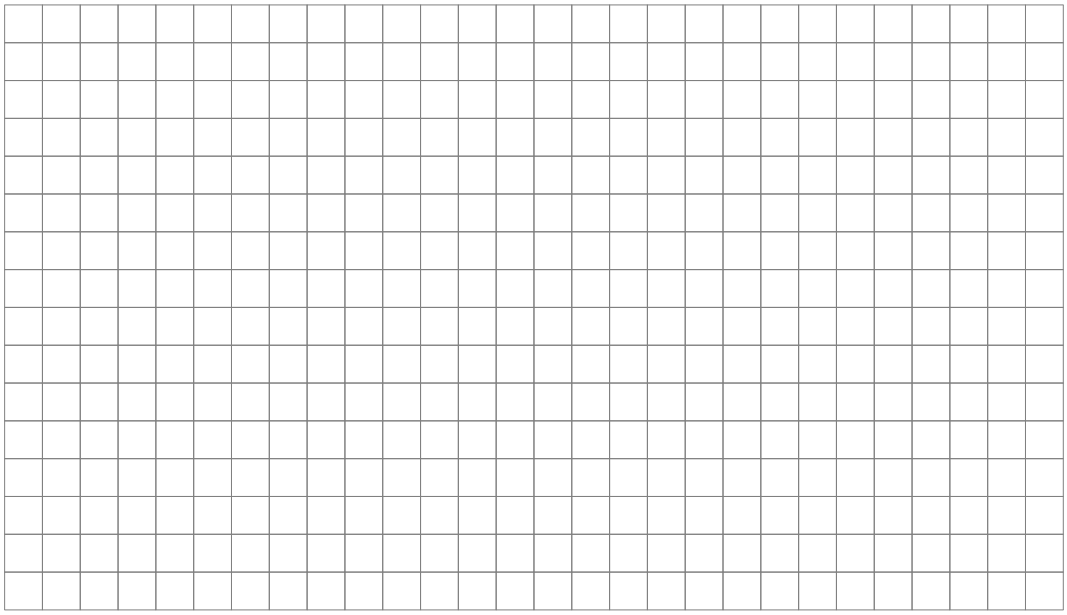
.....

.....

.....

Aufgabe 6: Data Analysis (6 Punkte)

6.1. Gegeben seien die Vektoren $\vec{v}_1 = (6, 9, 9, 4, 6)$ und $\vec{v}_2 = (4, 3, 3, 4, 1)$. Geben Sie die Euklidische und Manhattandistanz der beiden Vektoren an. (1)



d_{eukl}		d_{manh}	

6.2. Was ist der Unterschied zwischen Regression und Klassifizierung (*classification*)? Erklären Sie in höchstens vier Sätzen. (2)

6.3. Beschreiben Sie in höchstens 6 Sätzen den Ablauf einer Iteration des k-Means Algorithmus. (3)

Aufgabe 7: Multiple Choice (10 Punkte)

Diese Aufgabe umfasst 10 Multiple-Choice-Fragen. Sie bestehen aus jeweils mehreren Antwortmöglichkeiten, wobei mindestens eine Antwortmöglichkeit richtig und mindestens eine Antwortmöglichkeit falsch ist. Jede Frage, in der alle richtigen Antwortmöglichkeiten angekreuzt und alle falschen Antwortmöglichkeiten nicht angekreuzt sind, wird mit genau einem Punkt bewertet. Sollten nicht alle richtigen Antwortmöglichkeiten angekreuzt worden sein oder wurde mindestens eine falsche Antwortmöglichkeit angekreuzt, wird die Frage mit 0 Punkten bewertet.

7.1. Welche Aussagen gelten über die Armstrong-Axiome? (A,B und C seien beliebige Mengen (*sets*) funktionaler Abhängigkeiten (*functional dependencies*).) (1)

- ☐ Mit Hilfe der Armstrong-Axiome lassen sich aus einer Menge funktionaler Abhängigkeiten weitere funktionale Abhängigkeiten ableiten.
- ☐ Es gilt die Totalitätsregel (*rule of totality*): Es gilt entweder $A \to B$ oder $B \to A$, aber nicht beide.
- ☐ Es gilt die Symmetrieregel (*rule of symmetry*): Gilt $A \to B$, dann gilt auch $B \to A$.
- ☐ Es gilt die Transititivtätsregel (*rule of transitivity*): Gilt $A \to B$ und $B \to C$, dann gilt $A \to C$.

7.2. Welche Aussagen über referenzielle Integrität (*referential integrity*) gelten? (1)

- ☐ Referenzielle Integrität überprüft, ob alle Zeiger (*pointer*) im Programmcode der Datenbank initialisierten Speicher (*initialized memory*) referenzieren.
- ☐ Datenbanken kontrollieren die Einhaltung von referenzieller Integrität.
- ☐ Datenbanken können Verletzung von referenzieller Integrität durch DELETE- und UPDATE-Befehle automatisch korrigieren.
- ☐ Referenzielle Integrität gilt, wenn alle Fremdschlüsselattribute (*foreign key attributes*) nur Werte annehmen, die im referenzierten Primärschlüsselattribut (*primary key attribute*) angenommen werden.

7.3. Wählen Sie allem zum Ausdruck (*expression*) $\pi_a(\sigma_c(R))$ äquivalenten Ausdrücke, dabei sei a eine beliebige Attributmenge (*set of attributes*) und c ein beliebiges Selektionsprädikat. (1)

- ☐ $\sigma_c(\pi_a(R))$
- ☐ $R \cap \pi_a(\sigma_c(R))$
- ☐ $\sigma_c(R \bowtie_c R)$
- ☐ $\pi_a(\sigma_c(R) \cup \sigma_c(R))$

7.4. Wie viele Relationen werden benötigt, um eine partielle (*partial*) und disjunkte (*disjoint*) Generalisierung/Spezialisierungsbeziehung mit n Spezialisierungen im objektorientierten Stil abzubilden? (1)

- ☐ n
- ☐ 2^n
- ☐ $n + 1$
- ☐ 1

7.5. Gegeben seien die Relationen R und S sowie der Ausdruck (*expression*) $\pi_a(R \times S)$, dabei sei a eine beliebige Attributliste (*set of attributes*). Geben Sie alle korrekten Aussagen über die mögliche Kardinalität des Ausdrucksergebnisses an. (*Mengen – sets, Multimengen – multisets*) (1)

- ☐ Das Ergebnis enhtält nie mehr als $|R| \cdot |S|$ Tupel.
- ☐ Für Multimengen kann das Ergebnis mehr Tupel enthalten als für Mengen.
- ☐ Für Mengen kann das Ergebnis mehr Tupel enthalten als für Multimengen.
- ☐ Das Ergebnis kann nicht die leere Menge sein.

7.6. Welches der ACID-Kriterien beschreibt die folgende Aussage: „Ein Systemausfall darf nicht zum Verlust bereits erfolgreich abgeschlossener Transaktionen führen. (*Successfully completed transactions must not be undone by system failures.*)“ (1)

- ☐ Atomicity
- ☐ Consistency
- ☐ Isolation
- ☐ Durability

7.7. Gegeben sei das folgende XML-Dokument: (1)

```
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
  <food id="3">
    <name language="en-gb">Belgian Waffles</name>
    <price currency="USD">5.95</price>
    <description>
      Two of our famous Belgian Waffles with plenty of real maple syrup
    </description>
    <calories>650</calories>
  </food>
  <food id="5">
    <name language="en-us">Strawberry Belgian Waffles</name>
    <price currency="USD">7.95</price>
    <calories />
  </food>
</breakfast_menu>
```

Welche Aussagen sind korrekt? (*Knoten* - *node*)

- ☐ Das XML-Dokument ist syntaktisch korrekt.
- ☐ Es handelt sich um ein strukturiertes (*structured*) XML-Dokument.
- ☐ Der `currency`-Knoten ist ein Element-Knoten
- ☐ Das XML-Dokument besteht ausschließlich aus nicht-leeren (*non-empty*) Knoten.

7.8. Welche der folgenden Aussagen über den k-Means-Algorithmus sind korrekt? (1)

- ☐ Eine Möglichkeit zur Wahl der initialen Clusterzentren ist die Auswahl k zufälliger, unterschiedlicher Punkte aus der Datenmenge.
- ☐ Der k-means-Algorithmus wird verwendet um Strukturen in Daten zu erkennen.
- ☐ Der k-means-Algorithmus verwendet die Manhattan-Distanz als Distanzfunktion.
- ☐ Das Ergebnis des Algorithmus sind die Menge der finalen Clusterzentren sowie eine Zuordnung der Punkte zu den Clusterzentren.

7.9. Welche der folgenden Aussagen über den Reservoir Sampling-Algorithmus sind korrekt? (1)

- ☐ Die Anzahl der Elemente im Sample wächst unbegrenzt mit der Anzahl der gesehenen Elemente.
- ☐ Jedes im Datenstrom gesehene Element hat die selbe Wahrscheinlichkeit (*probability*) im Sample enthalten zu sein.
- ☐ Ein Element, das nur ein einziges Mal im Datenstrom gesehen wurde, kann mehr als ein Mal im Sample enthalten sein.
- ☐ Ein Element, das nur ein einziges Mal im Datenstrom gesehen wurde, kann höchstens ein Mal im Sample enthalten sein.

7.10. Für welche der folgenden Arten von Windows gilt die folgende Aussage: „Die Anzahl der Elemente im Fenster ist nach oben durch einen festen Wert begrenzt. (*The number of elements in the window has a fixed upper bound.*)“ (1)

- ☐ Time-based Sliding Window
- ☐ Tuple-based Sliding Window
- ☐ Time-based Landmark Window
- ☐ Tuple-based Landmark Window