+1/1/60+

Informationssysteme und Datenanalyse

Raum	
Platz	

Schriftlicher Test

13.07.2019

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

Hinweise:

- Die Bearbeitungszeit für diesen Test beträgt 60 Minuten plus 10 Minuten Einlesezeit. Es können in 7 Themen insgesamt 50 Punkte erreicht werden. Während der Einlesezeit darf **nicht** gekreuzt oder geschrieben werden.
- Dieser Test besteht aus **16** Seiten. Bitte überprüfen Sie die Vollständigkeit der Seiten direkt nach Beginn der Einlesezeit.
- Dieser Test beinhaltet zwei Fragetypen. Bei Fragen von Typ 1 ist genau eine Antwortmöglichkeit korrekt. Bei Fragen von Typ 2 sind entweder eine oder mehrere Antwortmöglichkeiten korrekt. Fragen von Typ 2 sind mit dem Symbol ♣ markiert.
- Bei Fragen von Typ 2 vergeben wir Teilpunkte, wenn Sie einen Teil der richtigen Antwortmöglichkeiten ankreuzen. Wenn Sie eine oder mehrere falsche Antwortmöglichkeit ankreuzen, erhalten Sie 0 Punkte für die Frage.
- 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	

Aufgabe	**Punkte**	**Erreicht**
EER-Modellierung	7	
Relationaler Entwurf	7	
Anfragesprachen	11	
Transaktionen	5	
Data Warehousing	3	
Data Streams Management	7	
Data Science	10	
Total	50	

(Erweiterte) Entity-Relationship-Modellierung

Das Sensornetzwerk

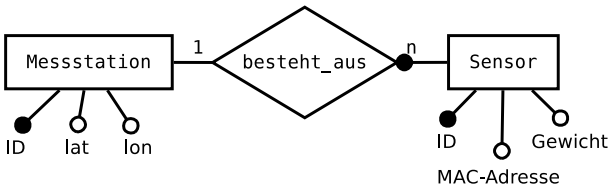
Im folgenden sind 7 (E)ER-Diagramme gegeben, von denen Sie beurteilen sollen, ob diese aus dem unten stehenden Text über eine Datenbank eines Sensornetzwerks abgeleitet werden können.

Hinweis: *Es empfiehlt sich, erkannte Entitytypen in den Texten zu markieren.*

Im Rahmen eines Forschungsprojektes soll ein Sensornetzwerk in einer Datenbank abgebildet werden. Eine Messstation mit eindeutiger ID, lat und lon besteht dabei immer aus verschiedenen Sensoren, welche nicht mit einer Messsattion verbunden sein müssen. Alle Sensoren besitzen eine eindeutige ID sowie eine MAC-Adresse, ein Gewicht kann zusätzlich auch gespeichert werden. Sensoren sind immer vom Typ „Temperatur“, „NOX“, „PM10“ oder „Hydro“, es können dabei auch die Mischformen „NOX“ und „PM10“ bzw. „Temperatur“ und „Hydro“ auftreten.

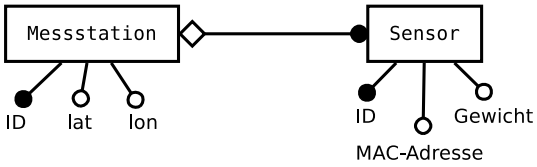
Frage 1 (1 Punkt)

☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



Frage 2 (1 Punkt)

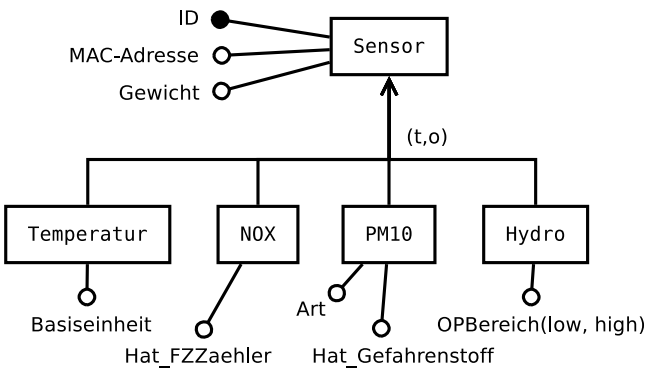
☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



Temperatursensoren besitzen darüber hinaus noch eine Basiseinheit, während NOX-Sensoren einen Fahrzeugzähler besitzen können. PM10-Sensoren sind auf eine bestimmte Weise hergestellt worden und dürfen auf Grund von möglicherweise enthaltenen Gefahrenstoffen teilweise nur von speziell ausgebildetem Personal gewartet werden. Hydro-Sensoren können nur in bestimmten Wertebereichen operieren. Darüber hinaus können Sensoren einen Hersteller haben, der sich über einen Namen sowie eine Postleitzahl identifizieren lassen und außerdem eine Bilanz-URL besitzen.

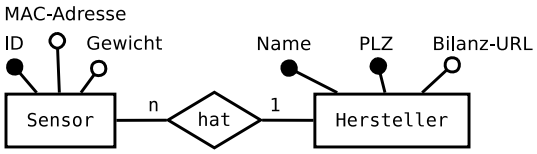
Frage 3 (1 Punkt)

☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



Frage 4 (1 Punkt)

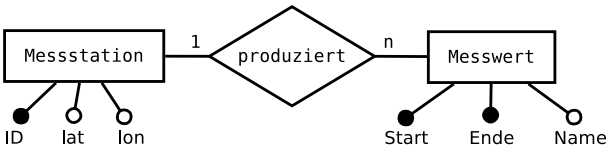
☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



Es existieren weiterhin Messzyklen mit einer eindeutigen Nummer sowie einer Mittelwertfunktion. Beliebig viele Messzyklen können aufeinander folgen. Messwerte müssen Teil eines solchen Messzyklus sein. Sie werden durch ein Konfidenzintervall, bestehend aus Start- und Endzeitpunkt identifiziert. Zusätzlich wird auch ein zufälliger Name gesetzt. Mehrere Messwerte können mit beliebig vielen Sensoren in Verbindung stehen, jeder Beziehung ist ein Zeitstempel zugeordnet.

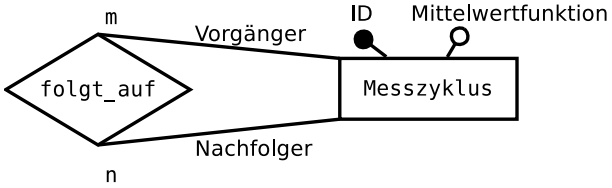
Frage 5 (1 Punkt)

☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



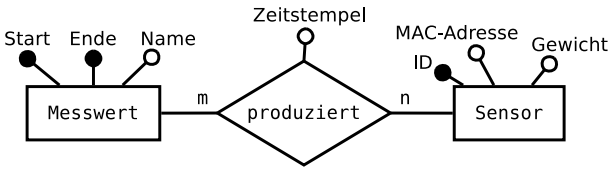
Frage 6 (1 Punkt)

☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



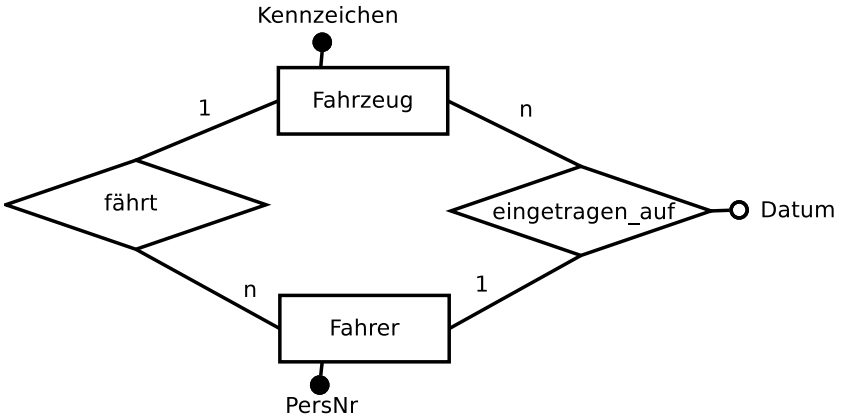
Frage 7 (1 Punkt)

☐ Ist abgebildet.
☐ Ist *nicht* abgebildet.



Das Relationale Modell

Frage 8 ♣ (2 Punkte) Welche der folgenden Abbildungen dieses ER-Diagramms in ein relationales Modell sind kapazitätserhaltend?



☐ Fahrer(PersNr), Fahrzeug(Kennzeichen),
fährt(PersNr → Fahrer, Kennzeichen → Fahrzeug),
eingetragen_auf(Persnr → Fahrer, Kennzeichen → Fahrzeug, , Datum)

☐ Fahrer(PersNr), Fahrzeug(Kennzeichen),
fährt(PersNr → Fahrer, Kennzeichen → Fahrzeug),
eingetragen_auf(PersNr → Fahrer, Kennzeichen → Fahrzeug, Datum)

☐ Fahrer(PersNr), Fahrzeug(Kennzeichen),
fährt(PersNr → Fahrer, Kennzeichen → Fahrzeug,)
eingetragen_auf(PersNr → Fahrer, Kennzeichen → Fahrzeug, Datum)

☐ Fahrer(PersNr, Kennzeichen → Fahrzeug, Datum),
Fahrzeug(Kennzeichen, PersNr → Fahrer)

☐ Fahrer(PersNr), Fahrzeug(Kennzeichen, Datum),
fährt(PersNr → Fahrer, Kennzeichen → Fahrzeug),
eingetragen_auf(PersNr → Fahrer, Kennzeichen → Fahrzeug)

☐ *Keine dieser Antworten ist korrekt.*

Frage 9 ♣ (1 Punkt) Gegeben sei ein EER Diagram mit einer Generalisierung/Spezialisierungsbeziehung zwischen einem generellen Entitytypen und 5 spezialisierten Typen. Die Beziehung ist überlappend und total. Wie viele Relationen werden zur Abbildung im objektorientierten Stil benötigt?

☐ 1
☐ 5
☐ 6
☐ 31
☐ 32
☐ *Keine dieser Antworten ist korrekt.*

Frage 10 ♣ (2 Punkte) Gegeben sei die Relation $R(A, B, C, D, E)$ sowie die folgenden funktionalen Abhängigkeiten. Welche der Attributmengen sind Superschlüssel?

$$\{B\} \to \{A, B\}, \{E, B\} \to \{C\}, \{C\} \to \{D\}$$

- ☐ $\{A, B, C, D, E\}$
- ☐ $\{A, C, D, E\}$
- ☐ $\{B\}$
- ☐ $\{B, C, D\}$
- ☐ $\{B, C, E\}$
- ☐ $\{B, E\}$
- ☐ $\{E\}$
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 11 ♣ (2 Punkte) Gegeben sei die Relation R($\underline{K_1, K_2}$, $\{S\}$, A_1, A_2, A_3) und die folgenden funktionalen Abhängigkeiten. Normalisieren Sie bis zur BCNF. Wie viele Relationen enthält der relationale Entwurf in BCNF?

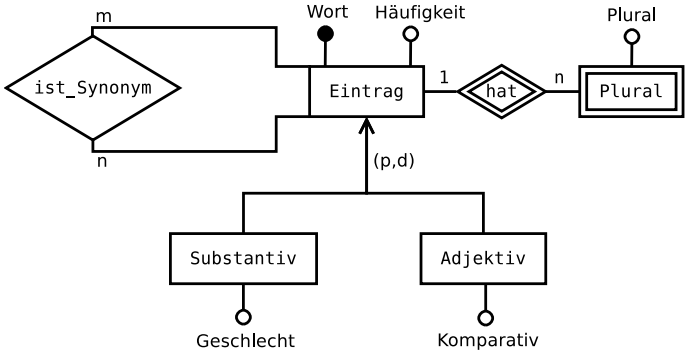
$$\{A_3\} \to \{K_1\}, \{A_2\} \to \{A_1\}$$

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7
- ☐ *Keine dieser Antworten ist korrekt.*

Anfragesprachen

Das Wörterbuch

Gegeben sei folgendes Schema eines *Wörterbuchs* mit Beispieltupeln.



Adjektiv

Wort	Häufigkeit	Komparativ
schön	4	schöner
ganz	5	`NULL`
elysisch	2	elysischer

Eintrag

Wort	Häufigkeit
doch	5
nicht	5
schon	5

Ist_Synonym

Wort	**Synonym**
Schule	Bildungsstätte
Schule	Lehranstalt
Lehranstalt	Schule
Bildungsstätte	Schule
Universität	Lehranstalt
Holler	Unsinn
Unsinn	Holler
schön	elysisch
elysisch	schön

Plural

Wort	**Plural**
Wort	Wörter
Wort	Worte
Holler	Holler
Bildungsstätte	Bildungsstätten
Schule	Schulen
Lehranstalt	Lehranstalten
Universität	Universitäten
elysisch	elysischen
schön	schönen

Substantiv

Wort	Häufigkeit	Geschlecht
Wort	4	N
Schule	4	F
Holler	1	M
Bildungsstätte	2	F
Unsinn	3	M
Lehranstalt	2	F
Universität	4	F

Frage 12 ♣ (2 Punkte) Welche der folgenden Anfragen sind äquivalent zu:

```
SELECT p.wort wort, COUNT(*) count
FROM plural p, substantiv s
WHERE p.wort = s.wort
GROUP BY p.wort
```

- ☐ $\gamma_{wort,COUNT(*)\rightarrow count}(\sigma_{wort=swort}(plural \times \rho_{S(swort)}(\pi_{wort}(substantiv))))$
- ☐ $\gamma_{wort,COUNT(*)\rightarrow count}(plural)$
- ☐ $\gamma_{wort,COUNT(*)\rightarrow count}(plural \bowtie substantiv)$
- ☐ $substantiv \bowtie_{wort=wort} \gamma_{wort,COUNT(*)\rightarrow count}(plural)$
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 13 (2 Punkte) Wie viele Tupel werden von der folgenden Anfrage ausgegeben:

```
SELECT *
FROM adjektiv a JOIN plural p ON a.wort = p.wort
WHERE 1 = (SELECT COUNT(*) FROM plural WHERE wort = a.wort)
```

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 9

Frage 14 (3 Punkte) Welche Aussage beschreibt die folgende Anfrage:

```
SELECT a.wort, a.synonym
FROM ist_synonym a, ist_synonym b
WHERE a.wort = b.synonym AND a.synonym = b.wort
GROUP BY a.wort, a.synonym
HAVING a.wort > a.synonym
```

Hinweis: *In SQL wird beim Vergleichen zweier Zeichenketten mit* `<` *und* `>` *anhand der lexikografischen Ordnung („alphabetisch“) verglichen. Beispiel:* `Baum < Bieber == TRUE; Apfel < Haus == TRUE`

- ☐ Eliminieren Sie alle synonymen Wortpaare, die symmetrisch sind (d.h., Wort a ist ein Synonym für Wort b und Wort b ist ein Synonym für Wort a). Es sollen nur asymmetrische Wortpaare ausgegeben werden.
- ☐ Finden Sie alle synonymen Wortpaare, die symmetrisch sind (d.h., Wort a ist ein Synonym für Wort b und Wort b ist ein Synonym für Wort a). Jedes dieser Wortpaare soll genau einmal ausgegeben werden.
- ☐ Finden Sie alle synonymen Wortpaare, wobei das Wort alphabetisch nach seinem Synonym gereiht ist.
- ☐ Gruppieren Sie alle Worte nach ihren jeweiligen Synonymen in alphabetisch aufsteigender Reihenfolge.

Frage 15 ♣ (3 Punkte) Welche der folgenden Anfragen erfüllt: *Finden Sie die durchschnittliche Häufigkeit über alle Worte.*

☐
```
SELECT (SUM(sum) / SUM(count)) häufigkeit
FROM (
    SELECT SUM(häufigkeit) sum, COUNT(*) count FROM eintrag
    UNION ALL SELECT SUM(häufigkeit) sum, COUNT(*) count FROM substantiv
    UNION ALL SELECT SUM(häufigkeit) sum, COUNT(*) count FROM adjektiv
) r
```

☐
```
SELECT AVG(häufigkeit) häufigkeit
FROM (
    SELECT häufigkeit FROM eintrag
    UNION ALL SELECT häufigkeit FROM substantiv
    UNION ALL SELECT häufigkeit FROM adjektiv
) r
```

☐
```
SELECT AVG(h) häufigkeit
FROM (
    SELECT AVG(häufigkeit) h FROM eintrag
    UNION SELECT AVG(häufigkeit) h FROM substantiv
    UNION SELECT AVG(häufigkeit) h FROM adjektiv
) r
```

☐
```
SELECT (se + ss + sa) / (ce + cs + ca) häufigkeit
FROM
    (SELECT SUM(häufigkeit) se, COUNT(*) ce FROM eintrag) e,
    (SELECT SUM(häufigkeit) ss, COUNT(*) cs FROM substantiv) s,
    (SELECT SUM(häufigkeit) sa, COUNT(*) ca FROM adjektiv) a
```

☐ *Keine dieser Antworten ist korrekt.*

Frage 16 (1 Punkt) Was ist das Ergebnis folgender Anfrage?

$\sigma_{wort='Schule'}(\pi_{wort}(ist\_synonym))$

☐ { }

☐ { (Schule) }

☐ { (Schule) , (Schule) }

☐ { (Schule, Bildungsstätte), (Schule, Lehranstalt) }

☐ *Keine dieser Antworten ist korrekt.*

Transaktionen

Frage 17 ♣ (2 Punkte) Welche Aussagen über Schedule S_2 sind korrekt?

$$S_2 = r_1(A)\ \ w_1(A)\ \ r_2(A)\ \ w_2(A)\ \ r_1(B)\ \ w_1(B)\ \ r_2(B)\ \ w_2(B)$$

- ☐ S_2 ist ein serieller Schedule
- ☐ S_2 ist konfliktserialisierbar
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 18 (1 Punkt) Welches Problem tritt bei dem folgenden Schedule auf?

Transaktion 1	Transaktion 2
$read(A, x)$	
$x = x + 15$	
	$read(A, y)$
	$y = y + 30$
$write(x, A)$	
	$write(y, A)$
commit	
	commit

- ☐ Non-repeatable Read
- ☐ Dirty Read
- ☐ Lost Update

Frage 19 ♣ (2 Punkte) Welche Aussagen über Schedule S_1 sind korrekt?

$$S_1 = r_1(A)\ \ r_1(B)\ \ r_2(A)\ \ w_2(A)\ \ r_2(B)\ \ w_1(A)\ \ w_1(B)\ \ w_2(B)$$

- ☐ S_1 ist ein serieller Schedule
- ☐ S_1 ist konfliktserialisierbar
- ☐ *Keine dieser Antworten ist korrekt.*

Data Warehousing

Frage 20 ♣ (1 Punkt) Welche der folgenden Aussagen zu einem Data Warehouse sind korrekt?

- ☐ Ein Data Warehouse ist eine OLAP-optimierte Datenbank, deren Inhalt aus verschiedenen anderen Datenquellen extrahiert wird.
- ☐ Ein Data Warehouse ist eine OLTP-optimierte Datenbank, deren Inhalt aus verschiedenen anderen Datenquellen extrahiert wird.
- ☐ Das Pentagrammschema ist eine mögliche relationale Repräsentation eines OLAP Würfels.
- ☐ Das Schneeflockenschema ist eine mögliche relationale Repräsentation eines OLAP Würfels.
- ☐ Das Fullfactschema ist die relationale Repräsentation eines OLAP Würfels mit dem geringsten Speicherverbrauch.
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 21 (1 Punkt) Wie viele Tabellen befinden sich in einem Schneeflockenschema für einen OLAP Würfel mit vier Dimensionen, die auf jeweils fünf Hierarchieebenen betrachtet werden können?

☐ 3	☐ 5	☐ 20	☐ 22
☐ 4	☐ 6	☐ 21	☐ 23

Frage 22 ♣ (1 Punkt) In welcher Normalform befindet sich ein beliebiges Fullfact-Schema in jedem Fall?

- ☐ 1. Normalform
- ☐ 2. Normalform
- ☐ 3. Normalform
- ☐ BCNFs Normalform
- ☐ *Keine dieser Antworten ist korrekt.*

Data Stream Management

Frage 23 ♣ (3 Punkte) Der Datenstrom (13, 14, 15) wird mit einem Bloomfilter mit 10 Bits und den Hashfunktionen $h_0(x)$ und $h_1(x)$ aufgezeichnet.

$$h_0(x) = ((x + 2) \bmod 15) \bmod 10$$
$$h_1(x) = ((2x) \bmod 12) \bmod 10$$

Welche der folgenden Aussagen über den Datenstrom sind anhand des Bloomfilters möglich?

- ☐ 11 kommt mindestens einmal vor.
- ☐ 11 kommt möglicherweise vor.
- ☐ 11 kommt nicht vor.
- ☐ 13 kommt mindestens einmal vor.
- ☐ 13 kommt möglicherweise vor.
- ☐ 13 kommt nicht vor.
- ☐ 43 kommt mindestens einmal vor.
- ☐ 43 kommt möglicherweise vor.
- ☐ 43 kommt nicht vor.
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 24 ♣ (2 Punkte) Eine Firma verkauft unter anderem drei Arten von Produkten: T-Shirts, Jeans und Jacken. Zwei Filialen der Firma nutzen jeweils einen Count-Min Sketch, um ihre Verkaufszahlen aufzuzeichnen. Beide Filialen verwenden die Hashfunktionen h_0 und h_1:

	h_0	h_1
T-Shirts	0	1
Jeans	1	0
Jacken	1	1
Schuhe	0	0

Die Sketches der jeweiligen Filiale sehen wie folgt aus.

h_0	7	15
h_1	8	10

Tabelle 1: Sketch der Filiale 1

h_0	8	13
h_1	9	7

Tabelle 2: Sketch der Filiale 2

Bezogen auf beide Filialen: Welche Verkaufszahlen sind mit diesen Sketches möglich?

- ☐ 17 Jeans
- ☐ 23 Jeans
- ☐ 24 Jeans
- ☐ 10 T-Shirts
- ☐ 14 T-Shirts
- ☐ 16 T-Shirts
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 25 ♣ (2 Punkte) Gegeben sind drei Hashfunktionen h_0, h_1 und h_2:

	h_0	h_1	h_2
Linh	0	2	1
Nils	0	0	0
Rudi	1	2	1
Bob	1	1	0
Anne	0	1	1

Welche Aussagen über einen Count-Min Sketch mit diesen Hashfunktionen sind korrekt?

- ☐ Linhs Kardinalität wird exakt wiedergegeben.
- ☐ Nils' Kardinalität wird exakt wiedergegeben.
- ☐ Bobs Kardinalität wird exakt wiedergegeben.
- ☐ *Keine dieser Antworten ist korrekt.*

Data Science

Frage 26 ♣ (2 Punkte) Welche Aussagen sind korrekt?

- ☐ *Leave-one-out*-Validierung benötigt weniger Berechnungen als Kreuzvalidierung.
- ☐ Es ist ein Zeichen für Overfitting, wenn der Vorhersagefehler auf den Testdaten kleiner ist als auf den Trainingsdaten.
- ☐ k-Means gehört zu den Algorithmen des überwachten Maschinellen Lernens.
- ☐ *Keine dieser Antworten ist korrekt.*

Frage 27 ♣ (2 Punkte)
Welche der folgenden Clusterings sind als Ergebnis einer konvergierten k-Means-Clusteranalyse **unmöglich**?

- ☐ Clustering A
- ☐ Clustering B
- ☐ Clustering C
- ☐ Clustering D
- ☐ *Keine dieser Antworten ist korrekt.*

Hierarchische Clusteranalyse

Wir nutzen hierarchisches Clusteranalyse mit euklidischer Distanzfunktion, um die Menge der natürlichen Zahlen von 1 bis 512 (inklusive 1 und 512) zu clustern.

Falls mehrere Cluster dieselbe Distanz aufweisen, werden die zwei Cluster zusammengeführt, die die kleinste Zahl beinhalten. Wenn zum Beipsiel Cluster A und B dieselbe Distanz zueinander haben wie Cluster C und D, führen wir A und B zusammen falls $\min(A \cup B) < \min(C \cup D)$. Falls $\min(A \cup B) = \min(C \cup D)$ entscheidet die nächstkleinere Zahl.

Wir interessieren uns dafür, wie groß die beiden Cluster sind, die wir als letztes zusammenführen (also die Cluster, die am Ursprung des Dendrogramms anliegen).

Frage 28 (2 Punkte) Wie groß sind die letzten beiden Cluster für single-linkage clustering?

- ☐ 256 und 256
- ☐ 257 und 255
- ☐ 510 und 2
- ☐ 511 und 1

Frage 29 (2 Punkte) Wie groß sind die letzten beiden Cluster für complete-linkage clustering?

- ☐ 256 und 256
- ☐ 257 und 255
- ☐ 510 und 2
- ☐ 511 und 1

Frage 30 (2 Punkte) Ein Datensatz wird mit folgender Funktion klassifiziert:

$$\hat{y}(\mathbf{x}) = \begin{cases} P, & \text{falls } f(\mathbf{x}) > 0 \\ Q, & \text{sonst} \end{cases}$$

mit $f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$. Die Klassifizierungsfunktion hat die Parameter $\mathbf{w} = (w_0, w_1, w_2)$. Jeder Datenpunkt ist ein Vektor $\mathbf{x} = (x_1, x_2)$, mit einer Klasse c. Der gesamte Datensatz ist:

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & c \end{array} \\ \begin{pmatrix} 1 & 1 & P \\ 1 & 2 & P \\ 3 & 2 & Q \\ 4 & 3 & Q \end{pmatrix} \end{array}$$

Was ist der Recall für die Parameterkombination $\mathbf{w} = (3, 1, -2)$ in diesem Klassifikationsmodell, wenn P für "positive" (1) steht?

- ☐ 0.0
- ☐ 0.33
- ☐ 0.40
- ☐ 0.50
- ☐ 0.60
- ☐ 0.67
- ☐ 0.9
- ☐ 1.0