MATEMATISKA INSTITUTIONEN
STOCKHOLMS UNIVERSITET
Avd. Beräkningsmatematik
Examinator: Lars Arvestad

Tentamensskrivning i
DA4001 Databasteknik
7.5 hp
2025-12-17

- Inga hjälpmedel tillåtna.
- **Skriv tydligt.** Svårlästa svar riskerar 0 poäng.
- Skriv bara på en sida av varje papper!
- Motivera alla svar (om inte annat anges)!
- **Betygsgränser:** E: 25, D: 30, C: 35, B: 40, A: 45

Note: the exam is available in English from page 4!

---

1. (a) Vad innebär *SQL-injektion* (eng: *SQL injection*). (2p)

(b) Förklara vad en *kandidatnyckel* är. (2p)

(c) Förklara vad en *entitetstyp* är. (2p)

(d) Förklara hur användning av `CREATE INDEX` kan göra dina databassökningar mer effektiva. (2p)

(e) Ange två anledningar till att inte skapa index för alla attribut i alla tabeller. (2p)

(f) Ge ett förväntat utdata av `SELECT * FROM Table1, Table2;` givet tabellerna i figur 1. (2p)

(g) Hur lägger man till en rad med värdena 3 och 7 till Table2? (2p)

Table1

Attr1	Attr2
1	3
2	4

Table2

ColA	ColB
1	5
2	6

Figur 1: Exempeldata för fråga 1f och 1g.

2. I figur 2 återges ett förenklat databasschema för proteindatabasen du använde i laboration 2. Använd den för att lösa följande uppgifter med SQL.

(a) Hur många proteinfamiljer finns definierade i databasen? (2p)

(b) Lista alla proteinfamiljer tillsammans med hur många proteiner de är tilldelade. Du kan anta att varje familj har minst ett protein. (2p)

(c) Lista den eller de proteinfamiljer som har flest ”medlemmar”. (2p)

(d) Lista proteiner som finns i samma familjer som ”STF1_HUMAN”. Observera att ett protein kan tillhöra flera familjer[1]. (2p)

(e) Säg att ett protein p_i är *länkat* till protein p_j om

- p_i och p_j ligger i samma familj, eller
- p_j ligger i samma familj som p_k som är länkat till p_i.

Lista alla proteiner som är länkade till STF1_HUMAN. (3p)

För att förtydliga tar vi ett exempel: om p_1 och p_2 ligger i familj X så är de länkade. Om p_2 också tillhör familj Y, där vi också finner p_3, så är p_1 också länkad med p_3. Om p_4 också ligger i Y samt i Z, så blir även p_4 länkad med p_1. Proteinet q som ligger i familj A, som inte är länkat med något protein i X, Y, eller Z, är dock inte länkat med p_1.

Ur ett tillämpat perspektiv kan man tolka frågan så här: vilka proteiner kan ha samma funktion som STF1_HUMAN om man antar att medlemskap i samma familj indikerar att proteinerna har samma funktion, och att begreppet ”funktion” är transitivt över familjer.

[1] Detta antagande skiljer sig mot vad som gällde på laborationen.

3. (a) Vad innebär ”rätten till radering av personuppgifter” enligt dataskyddsförordningen? (1p)

(b) Ge ett exempel på när man *inte* har rätt att få sina personuppgifter raderade. (1p)

(c) Ge två exempel på känsliga personuppgifter. (2p)

(d) Vad säger dataskyddsförordningen om *automatiserade beslut*? (1p)

4. Bo Taniker älskar växter och har mängder av dem. Han vill särskilt hålla koll på de krukväxter han har i sitt hus och i sina växthus. Bo gör noggranna mätningar av växterna, men är trött på att samla mätningarna i anteckningsböcker och ber dig därför om hjälp att skapa en relationsdatabas för ändamålet. Bo har hjälpsamt skapat och kommenterat ett ER-diagram som förklarar vad han vill ha, se figur 3.

(a) Vad innebär en romb i ett ER-diagram, som den i figurens mitt? (1p)

(b) Betrakta figur 3 och föreslå lämpliga relationer som noggrant följer Bos ER-diagram. (2p)

(c) ER-diagrammet är inte perfekt och relationerna kommer inte vara 3NF. Motivera varför. (3p)

(d) Föreslå en normalisering av relationerna till 3NF och motivera dina beslut. (4p)

Om du finner att förklaringarna från Bo i figur 3 inte ger dig all information du behöver så kan du förklara vad du saknar och föreslå en rimlig tolkning som är i linje med Bos behov.

Om du inte minns vad en relation är så kan du föreslå enkla tabeller istället.

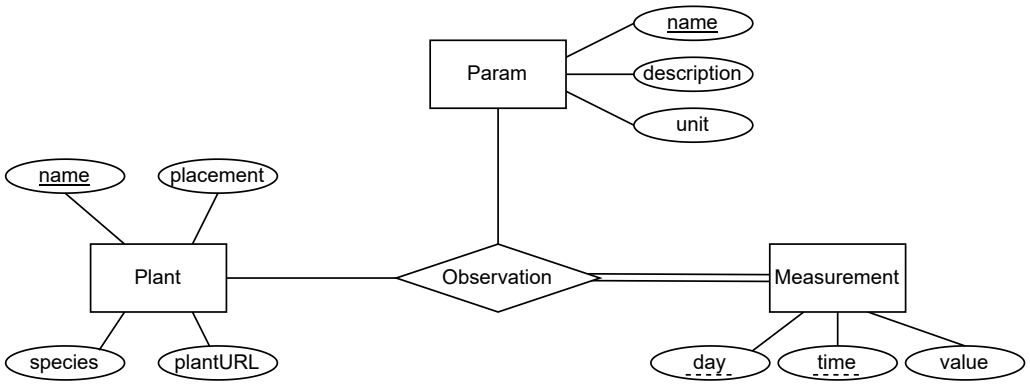
Du behöver inte rita ett nytt ER-diagram.

5. Skapa ett fullständigt databasschema baserat på dina valda relationer i fråga 4a eller 4c (ange vilket).

- Schemat ska ges som SQL och följa de relationer du föreslagit. (2p)
- Attributen ska ha typer lämpliga för MariaDB eller liknande. (2p)
- Primärnycklar ska framgå. (2p)
- Lämpliga referensvillkor ska finnas med. (2p)
- Ge minst ett exempel på ett semantiskt integritetsvillkor. (2p)

```
CREATE TABLE protein (
    accession   TEXT, -- Unik identifierare för protein
    species_id  INT,  -- NCBI taxonomic code
    mass        FLOAT,
    description TEXT,
    seq         TEXT,
);

CREATE TABLE protein_keywords (
    accession TEXT,
    keyword   TEXT
);

CREATE TABLE species (
    species_id INT,     -- NCBI taxonomic code
    abbrev TEXT,
    latin  TEXT,
    common TEXT
);

CREATE TABLE family (
    fam_acc TEXT, -- Unikt namn på proteinfamilj
    descr   TEXT  -- Vetenskaplig beskrivning av proteinfamilj
);

CREATE TABLE familymembers (
    family TEXT,  -- Refererar till fam_acc i tabellen "family"
    protein TEXT  -- Refererar till accession i tabellen "protein"
);
```

Figur 2: Förenklat databasschema för proteindatabasen från laboration 2.



Figur 3: **Ett ER-diagram för Bo Tanikers samling av växter.** Bo har unika och kärleksfulla namn (*plantname*) på sina växter och håller koll på var växterna står (*placement*, tex ”i köket” och ”östra växthuset”). Naturligtvis anger han vilken art en växt är (*species*) och skickar med en länk till motsvarande webbsida på PlantPedia (*plantURL*). Bo gör noggranna mätningar av sina växter och har flera parametrar som han följer. Varje parameter har ett namn (*paramname*, tex ”höjd”), en beskrivning (*description*, ”från jord till topp”) och vilken enhet han mäter i (*unit*, som ofta är ”cm”). En mätning görs vid ett visst datum (*day*) och tid (t).

The exam in English

- No aids allowed.
- **Write clearly.** Illegible answers risk 0 points.
- Write only on one side of each sheet of paper!
- Provide justification for all answers (unless otherwise stated)!
- **Grade boundaries:** E: 25, D: 30, C: 35, B: 40, A: 45

---

1. (a) Explain what *SQL injection* is. (2p)

(b) Explain what a *candidate key* is. (2p)

(c) Explain what an *entity type* is. (2p)

(d) Explain how the use of `CREATE INDEX` can make your database searches more efficient. (2p)

(e) Give two reasons why you should not create indexes for all attributes in all tables. (2p)

(f) Provide the expected output of `SELECT * FROM Table1, Table2;` given the tables in Figure 4. (2p)

(g) How do you insert a row with the values 3 and 7 into Table2? (2p)

Table1

Attr1	Attr2
1	3
2	4

Table2

ColA	ColB
1	5
2	6

Figur 4: Example data for question 1f and 1g.

2. Figure 5 shows a simplified database schema for the protein database you used in Lab 2. Use it to solve the following tasks with SQL.

(a) How many protein families are defined in the database? (2p)

(b) List all protein families together with how many proteins are assigned to them. You may assume that each family has at least one protein. (2p)

(c) List the protein family or families that have the most "members". (2p)

(d) List proteins that are in the same families as "STF1_HUMAN". Note that a protein may belong to several families[2]. (2p)

(e) Say that a protein p_i is *linked* to protein p_j if

- p_i and p_j are in the same family, or
- p_j is in the same family as p_k which is linked to p_i.

List all proteins that are linked to STF1_HUMAN. (3p)

To clarify, let us take an example: if p_1 and p_2 are in family X then they are linked. If p_2 also belongs to family Y, where we also find p_3, then p_1 is also linked with p_3. If p_4 also belongs to Y as well as Z, then p_4 also becomes linked with p_1. Protein q which belongs to family A, which is not linked with any protein in X, Y, or Z, is however not linked with p_1.

From an applied perspective, one can interpret the question like this: which proteins may have the same function as STF1_HUMAN if we assume that membership in the same family indicates that the proteins have the same function, and that the concept of "function" is transitive across families.

[2]This assumption differs from what applied in the lab.

3. (a) What does the ”right to erasure of personal data” mean according to the GDPR? (1p)

(b) Give an example of when one *does not* have the right to have their personal data erased. (1p)

(c) Give two examples of sensitive personal data. (2p)

(d) What does the GDPR say about *automated decisions*? (1p)

4. Bo Taniker[3] loves plants and has plenty of them. He especially wants to keep track of the potted plants he has in his house and in his greenhouses. Bo makes careful measurements of the plants, but is tired of collecting the measurements in notebooks and therefore asks you for help to create a relational database for this purpose. Bo has helpfully created and commented on an ER diagram that explains what he wants, see Figure 6.

(a) What does a rhombus in an ER diagram represent, like the one in the middle of the figure? (1p)

(b) Consider Figure 6 and propose suitable relations that closely follow Bo’s ER diagram. (2p)

(c) The ER diagram is not perfect and the relations will not be in 3NF. Motivate why. (3p)

(d) Propose a normalization of the relations to 3NF and motivate your decisions. (4p)

If you find that Bo’s explanations in Figure 6 do not give you all the information you need, you may explain what you are missing and propose a reasonable interpretation that is in line with Bo’s needs.

If you do not remember what a relation is, you may instead propose simple tables.

You do not need to draw a new ER diagram.

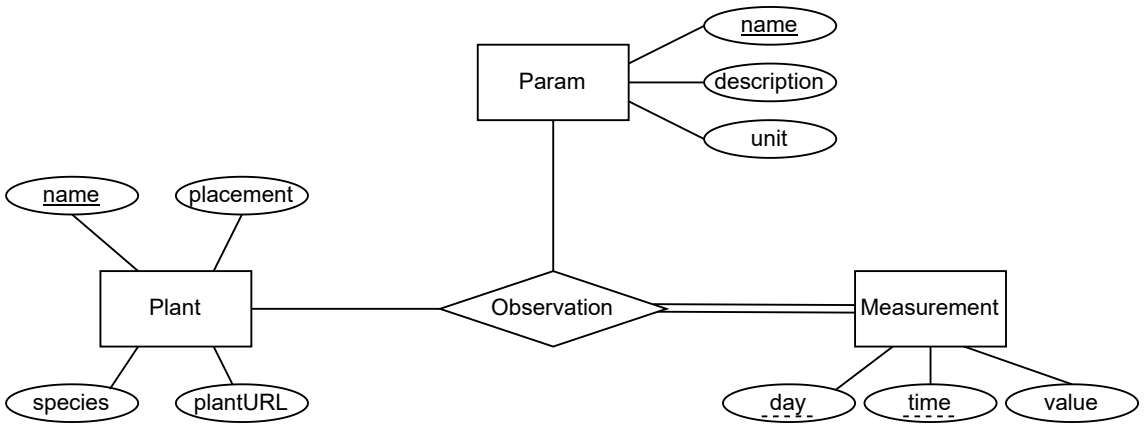
5. Create a complete database schema based on your chosen relations in question 4a or 4c (specify which).

- The schema should be given as SQL and follow the relations you have proposed. (2p)
- The attributes should have types suitable for MariaDB or similar. (2p)
- Primary keys should be specified. (2p)
- Appropriate referential constraints should be included. (2p)
- Provide at least one example of a semantic integrity constraint. (2p)

[3]This is a pun, combining a plausible Swedish name with ”botanist”.

```
CREATE TABLE protein (
    accession   TEXT, -- Unique identifier for protein
    species_id  INT,  -- NCBI taxonomic code
    mass        FLOAT,
    description TEXT,
    seq         TEXT,
);

CREATE TABLE protein_keywords (
    accession TEXT,
    keyword   TEXT
);

CREATE TABLE species (
    species_id INT,     -- NCBI taxonomic code
    abbrev TEXT,
    latin  TEXT,
    common TEXT
);

CREATE TABLE family (
    fam_acc TEXT, -- Unique name of protein family
    descr   TEXT  -- Scientific description of protein family
);

CREATE TABLE familymembers (
    family TEXT,  -- Refers to fam_acc in the "family" table
    protein TEXT  -- Refers to accession in the "protein" table
);
```

Figur 5: Simplified database schema for the protein database used in lab 2.



Figur 6: **An ER diagram for Bo Taniker's plant collection.** Bo has unique and affectionate names (*plantname*) for his plants and keeps track of where the plants are located (*placement*, e.g. "in the kitchen" and "eastern greenhouse"). Naturally, he specifies which species a plant belongs to (*species*) and provides a link to the corresponding webpage on PlantPedia (*plantURL*). Bo makes careful measurements of his plants and follows several parameters. Each parameter has a name (*paramname*, e.g. "height"), a description (*description*, "from soil to top"), and the unit in which he measures it (*unit*, often "cm"). A measurement is taken at a specific date (*day*) and time (*t*).