MATEMATISKA INSTITUTIONEN
STOCKHOLMS UNIVERSITET
Avd. Beräkningsmatematik
Examinator: Lars Arvestad

Tentamensskrivning i
DA4001 Databasteknik
7.5 hp
2026-02-11

- 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)! Det kan kännas onödigt för tex SQL-frågor, men det kan hjälpa rättande lärare att förstå vad du menar om det finns ett missförstånd.
- **Betygsgränser:** E: 25, D: 30, C: 35, B: 40, A: 45

---

1. Figur 1 visar ett databasschema för en enkel databas för ett tänkt marknadsföringsföretag. Figurtexten försöker förtydliga vad databasen lagrar.

Formulera följande frågeställningar i SQL, givet databasschemat i figur 1.

(a) Vad är den högsta registrerade årsinkomsten? (1p)

(b) Lista alla personer i databasen med förnamn, efternamn, och årsinkomst. (2p)

(c) Sammanställ hur många intressen vi har registrerade i de olika postnummerområdena. (2p)

Du kan ignorera de postnummerområden som saknar registrerade intressen. Utdata av ska vara en enkel tabell med kolumnerna ”Postnummer” och ”Antal intressen”, men du behöver inte ha de rubrikerna.

(d) Vad är den årliga lönen, i genomsnitt, för de som är intresserade av golf? Svara med hjälp av ett SQL-uttryck. (3p)

(e) Skapa en vy ”shared_interest” som är den delmängd av tabellen ”friendship” där vänner har ett delat intresse. Vänskapsrelationen ska ha minst ha 0,5 i ”styrka”. (3p)

(f) Skapa en lista med personer och deras intressen. Du ska visa förnamn, efternamn, och en beskrivning av personens intressen, om de finns. (2p)

Om en person har flera intressen registrerad så ska personen omnämnas flera gånger. Varje person ska finnas med minst en gång, även utan intressen. Ett möjligt utdata skulle vara:

```
Ali     Ens      Datorspel
Bo      Taniker  Växter
Bo      Taniker  Fjärilar
Cia     Ociao
Donna   Void     Programmering
```

2. Förbättra schemat i figur 1.

(a) Vilka referensvillkor kan man komplettera schemat i figur 1 med? (2p)

(b) Föreslå ett integritetsvillkor som går att uttrycka med SQL. (2p)

(c) Föreslå ett semantiskt integritetsvillkor som går att uttrycka med SQL. (2p)

3. Konstruera ett ER-diagram som överensstämmer med databasschemat i figur 1. Du ska använda den grafiska notation som kursboken har föreslagit. (5p)

4.

(a) Diskutera normaliseringen av databasschemat i fråga 1 utifrån Boyce-Codds normalform (BCNF). Vilka relationer når BCNF och hur motiverar du det? Om en relation inte når BCNF, varför inte? (3p)

(b) Skapa ett exempel, baserat på det givna schemat, som inte följer BCNF. Motivera varför. (2p)

5. Diskutera databasen i figur 1 med utgångspunkt i Dataskyddsförordningen. (5p)

6. Vad innebär objektifiering av sambandstyp? (2p)

7. Gör ett ER-diagram för följande verksamhet. Du ska använda den grafiska notation som kursboken har föreslagit. (10p)

En medicinsk forskningsorganisation, *AlbanoMed*, arbetar mycket med *cellinjer*, dvs odlingar av celler som man utför experiment på. Ett sådant exempel är de så kallade HeLa-cellerna som än idag odlas trots att de togs (utan tillstånd!) 1951 från en tumör hos Henrietta Lacks och blev den första lyckade odlingen *in vitro* (ungefär ”i provrör”) av mänskliga celler. [1]

För att skaffa en bättre överblick över cellinjer inom forskningen vill forskningsorganisationen strukturera och samla information om dem i en databas. Det är ditt uppdrag att hjälpa AlbanoMed. Det finns cellinjer från många olika modelldjur (dvs djur lämpliga för vetenskapliga studier) och för varje modelldjur finns det flera cellinjer. Vi måste ange vilket modelldjur cellinjen kommer ifrån.

För en cellinje kan man vilja hänvisa till en eller flera forskningsartiklar givna som ett PMID (en unik identifierare på sajten PubMed, givet som ett heltal) och laboratorier som man vet odlar eller har odlat cellinjen. Ett laboratorium kan ha kontaktinformation (föreslå en lämplig form) för varje cellinje. En del cellinjer har fått sin arvsmassan (dess DNA) fullständigt sekvenserad, ibland till och med flera gånger, och för dessa fall vill vi kunna hänvisa till en extern datakälla med en URL som anger var arvsmassan kan hämtas som datafil.

En ytterligare komplikation man vill hålla koll på är att en del cellinjer är kontaminerade av andra cellinjer. Det mest berömda fallet är just HeLa-celler som tack vare extraordinär härdighet överlever och sprids i laboratorier, och har kontaminerat flera andra cellinjer. En cellinje kan kontaminera flera andra cellinjer och en cellinje kan vara kontaminerad av flera olika cellinjer.

Hur strukturerar man bäst denna information i ett ER-diagram?

8. Föreslå lämpliga relationer (motsvarande tabeller) som överensstämmer med din ER-modell i fråga 7. Du behöver inte ange typer, primärnycklar, eller integritetsvillkor. (4p)

[1] Henrietta Lacks arvsmassa sprids alltså än idag i laboratorier över världen, men i degenererad form eftersom cancercellers arvsmassa är en ”trasig” version av arvsmassan.

```
CREATE TABLE anonymization (
       userid CHAR(10) PRIMARY KEY,
       fname  VARCHAR(50),
       lname  VARCHAR(50),
       );

CREATE TABLE characteristics (
       anonymous_user CHAR(10) PRIMARY KEY,
       gender         INTEGER,    -- 0=unknown, 1=man,2=woman,3=other
       adress_area    CHAR(6),    -- postnummer, area code
       education      VARCHAR(100),
       yearly_income  INTEGER     -- from Skatteverket
       );

CREATE TABLE user_interests (
       anonymous_user CHAR(10),
       interest_key   INTEGER,
       PRIMARY KEY(anonymous_user, interest_key)
       );

CREATE TABLE interests (
       interest_key  INTEGER PRIMARY KEY,
       interest_desc VARCHAR(100),
       yearly_cost   INTEGER
       );

CREATE TABLE friendship (
       user1    INTEGER,
       user2    INTEGER,
       strength FLOAT,            -- Estimated probability
       PRIMARY KEY(user1, user2)
       );
```

Figur 1: Möjlig primitiv databas för ett marknadsföringssystem. Vi håller identiteten på personer i en separat tabell (`anonymization`). Användar-id (`userid`) genererar vi själva. I de andra tabellerna använder vi attributet `anonymous_user` vilket refererar till `userid`. Den grundläggande klientinformation vi vill vill spara är egenrapporterad, förutom taxerad inkomst (`yearly_income`) som är från Skatteverket via våra marknadsföringspartners. Utbildningen läggs in i fritext, typ, men följer ett standardiserat skrivsätt så att vi kan hitta folk med samma utbildning och utbildningsnivå, tex ”Kandidatexamen i Matematik”. Tabellen `interests` knyter en hobby eller aktivitet till en unik identifierare (som knyts till användare i tabellen `user_interests`) och vår uppskattning av typisk årlig kostnad för aktiviteten. Slutligen har vi tabellen `friendship` som innehåller information från partners (FB, Twitter, mfl) där det är relativt sannolikt (bedömt i parametern `strength`) att två personer känner varandra. Det ska vara sannolikhet.