1 (16 points) Equivalences

For each of the following pairs of queries (in relational algebra or SQL), you will write the contents of two database instances. The databases have the following schemas:

```
A(a int, b int);
B(a int, b int)
```

(2 Points) The first database instance should be populated with one or more rows so that Q1 and Q2 output different results. If Q1 and Q2 are equivalent, then write "identical" next to the empty tables instead.

(2 Points) The second database instance should be populated with one or more rows so that Q1 and Q2 return the same results. If this is not possible, write "not possible" next to the empty tables instead.

1.1 (4 Points, 2 Per Database Instance)

Q1: $A \bowtie_a B$
Q2: $A \bowtie B$

1.2 (4 Points, 2 Per Database Instance)

Q1: $A \bowtie_a B$
Q2: $A \times \sigma_{B.a='a'}(B)$

1.3 (4 Points, 2 Per Database Instance)

Q1: $\sigma_{\$1=\$3}(A \times B)$
Q2: `SELECT * FROM A, B WHERE A.a = B.a`

1.4 (4 Points, 2 Per Database Instance)

Q1: `SELECT * FROM A JOIN B ON A.a = B.a WHERE B.a = 1 or B.b = 2`
Q2: `SELECT * FROM A JOIN B ON A.a = B.a WHERE B.a = 1`

```
 UNION ALL
SELECT * FROM A JOIN B ON A.a = B.a WHERE B.b = 2
```

2 (18 points) Entity-Relationship Models

2.1 Constraints (3 Points)

Your friend downloaded the following CSV file and shared it with you. What constraints, if any, can be inferred from this dataset? If there are none, simply write ”none” and explain why in one sentence:

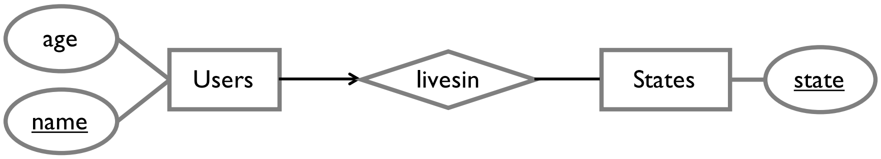
name	age	state
amy	18	TX
amy	18	CA
amy	18	FL
joe	20	MA
joe	20	NY

Table 1: Table for Problem 2

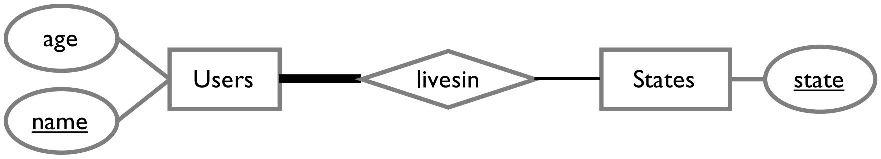
2.2 ER Constraints

For each of the following ER diagrams, select TRUE if Table 1 satisfies the constraints depicted in the diagram, and FALSE otherwise. If FALSE, write a short sentence about why.

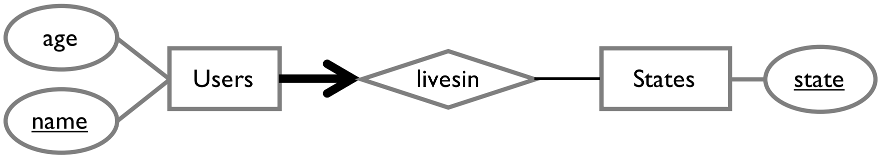
2.2.1 (3 Points)



2.2.2 (3 Points)



2.2.3 (3 Points)



2.3 ER to SQL

2.3.1 (3 Points)

Translate the ER diagram from Problem 2.2.1 into SQL.

2.3.2 (3 Points)

Translate the ER diagram from Problem 2.2.3 into SQL.

3 (8 points) Triggers

Consider the following database schema, and each table is empty:

```
CREATE TABLE A(a int);
CREATE TABLE B(b int);
```

In this question, we will present two implementations of triggers. Assuming the two tables are initially empty, and you run the following `INSERT` statements, you will write the final contents of `A` and `B` in the answer sheet.

```
INSERT INTO a VALUES(1);
INSERT INTO a VALUES(2);
INSERT INTO a VALUES(3);
```

3.1 Triggers

3.1.1 (4 Points, 2/table)

```
CREATE FUNCTION UDF1() RETURNS trigger
AS $$
BEGIN
  INSERT INTO b VALUES(NEW.*);
  RETURN NEW;
END;
$$ language plpgsql;
```

```
CREATE TRIGGER T1
BEFORE INSERT ON a
FOR EACH ROW
  EXECUTE PROCEDURE UDF1();
```

3.1.2 (4 Points, 2/table)

```
CREATE FUNCTION UDF2() RETURNS trigger
AS $$
BEGIN
  INSERT INTO b VALUES(NEW.*);
  RETURN null;
END;
$$ language plpgsql;
```

```
CREATE TRIGGER T2
AFTER INSERT ON a
FOR EACH ROW
  EXECUTE PROCEDURE UDF2();
```

4 (10 points) Misc. Questions

4.1 (2 Points)

In at most 2 short sentences, describe the significance of integrity constraints in database management systems as compared to writing code to check constraints within the application.

4.2 (2 Points)

List 2 important properties that the relational model provides that the Network/Hierarchical model does not provide. 4 words MAX for each property.

4.3 (2 Points)

In at most 2 short sentences, describe the difference between `VIEW` and `WITH` in SQL.

4.4 (2 Points)

In ONE sentence, explain multiset semantics.

4.5 (2 Points)

Write a creative example of joins In Real Life by filling in the sentence in the answer sheet. Most creative answer (subjectively judged by the staff) gets 2 extra credit points.

5 (14 points) Pass the SQL

The Warriors is the dominant basketball team in the National Basketball Association (NBA). Legend states that their dominance is not due to having a team of four (now five) NBA all stars, but instead due to their focus on passing and unselfish ball handling. Is this really the case? This problem will walk through an analysis to study how long players hold the ball before they pass to their teammates. When a player holds the ball, we say that the player *possesses* the ball.

Consider the following database schema, where Players contains information about each basketball player, Teams contains information about each team, and Possessions contains information about each time a player held the ball and how long the player possessed the ball.

```
Players(                  Teams(                       Possessions(
  pid int primary key,      tid int primary key,         id int primary key,
  tid int not null          name text not null,          pid int not null
      references Teams,     westcoast bool not null          references Players,
  name text not null,     )                              -- when the possession started
  age int not null                                       time timestamp not null,
                                                         -- number of seconds the player
)                                                        -- held the ball
                                                         held int not null
                                                       )
```

5.1 Sports Never Ages (2 Points)

Write the SQL query to find the average age across all players on west coast teams (westcoast is true). The output should be the average age.

5.2 Hold It (4 Points)

Fill in the `CREATE TABLE TimeHeld(pid, name, held` statement by writing a query that computes the average possession time for each player. The average possession time is defined as the average number of seconds that a player possesses the ball before the ball is passed to a teammate. Return the `pid` and `name` of each player, and the player's average possession time.

5.3 Teams That Pass (4 Points)

Fill in the `CREATE TABLE TeamPasses(passer, passee)` statement so it contains the `pid`s of the players that passed the ball (`passer`) and the teammate that received the pass (`passee`).

This is defined as `Possessions` where the passer that is in possession of the ball is on the same team as the passee that receives the ball, *and* if the time of the passee's possession is equal to the time of the passer's possession plus the amount of time that the passer held onto the ball. You may assume that `TimeHeld` has been correctly created and may use it in your answer if it helps.

5.4 The Control Tower (6 Points)

Fill in the `CREATE TABLE Control(tid, pid)` statement to identify the players on each team that have passed the ball to every teammate at least once. Return the team `tid` and player `pid`. You may assume that `TimeHeld` and `TeamPasses` have been correctly created and may use it in your answer if it helps.

5.5 The Passing-est (4 Points)

Let the team possession time be the average of the average possession time over all players on the team. Write the SQL query that returns the name of the team with the lowest team possession time. You may assume that `TimeHeld`, `TeamPasses`, and `Control` have been correctly created and may use it in your answer if it helps.

1 (10 points) Terms and Definitions

(2 points each) In *at most* two short sentences each, explain the meaning of the following terms as they relate to database management systems.

1. **Entity Set**

2. **Super Key**

3. **Null**

4. **Integrity Constraint**

5. **Natural Join**

2 (10 points) EJBank Relational Algebra

Evan's bank stores its data in two relations, with the following SQL schema:

```
CREATE TABLE Customers(
  cid int PRIMARY KEY,
  name text,
  state text
);

CREATE TABLE Accounts(
  aid int PRIMARY KEY,
  cid int NOT NULL REFERENCES cid,
  balance real NOT NULL
);
```

1. **(2 points)** Write a relational algebra expression to compute the account ids that have balances greater than $50,000.

2. **(4 points)** Write a relational algebra expression to compute the names of customers with balances greater than $50,000 in the state of "NY".

3. **(4 points)** Given the following values for the Account relation:

aid	cid	balance
1	102	1000.00
2	102	2000.00
3	107	2000.00
4	108	1000.00

What is the result of the following relational algebra expression?

$\rho(A, Accounts)$
$\rho(B, Accounts)$
$\pi_{A.aid,B.aid}(A \bowtie_{A.balance>B.balance} B)$

(continued on next page)

Fill in your answer in this table. Do not fill in the names for the fields. *Note*: you may or may not need all the columns and rows.

3 (20 points) Medical testing Entity-Relationship Modelling

A medical lab testing company has several testing centers all over the country. In this problem, you will design a schema to keep track of the testing centers, tests and order information. Specifically, you will need to keep track of:

1. The address, city, state and manager for each testing center.
2. The equipment at each testing center including the name of the machine, manufacturing year, and status of each machine (either functional or in repair). Note that one center may have multiple machines of the same kind.
3. The types of tests the company can perform. Tests have a name, time to run and price. The price of the test depends on the specific testing center. Some tests are only available at specific centers.
4. When an order for a test is submitted, the lab must record the doctor's name, the patient's name, and the date it was ordered. Each order is performed at a specific testing center. It must be performed at a center that offers that type of test.

Part 1: (10 points) Draw an ER diagram representing your database. Include 1-3 sentences of justification for why you drew it the way you did.

Part 2: (10 points) Write a SQL schema for your database. Include 1-3 sentences of justification for why you chose the tables you did.

(*Additional space on the following page*)

4 (20 points) Wikipedia in SQL

For this question, we will use a simplified schema based on Wikipedia, shown below. Each page has a unique id, a human readable title, and the length of the page (in bytes). Links between pages are stored in the Link table. The source is the page that contains the link, and dest is the page that the link points to. As an example, a Link tuple with values (source=50, dest=100) means that page id 50 contains a link to page id 100.

```
CREATE TABLE Page(
  id int PRIMARY KEY,
  title text NOT NULL,
  length int NOT NULL
);

CREATE TABLE Link(
  source int REFERENCES Page,
  dest int REFERENCES Page,
  PRIMARY KEY (source, dest)
);
```

1. **(6 points)** Circle true or false for the following statements:

   (a) **True / False** There can be two pages with the title “Venus”.

   (b) **True / False** Broken links may exist (a link where the source or the destination pages do not exist).

   (c) **True / False** A page can only be the source of one link.

   (d) **True / False** If the page “Venus” contains a link to “Mars”, deleting “Venus” will not be permitted.

   (e) **True / False** If the page “Venus” is not the source of any links, deleting it will be permitted.

   (f) **True / False** Renaming pages is not permitted.

   Write SQL queries to answer the following questions:

2. **(2 points)** What is the id and title of all pages that have titles that begin with the string “Database”?

3. **(2 points)** What are the titles of the 10 longest pages (in bytes)?

4. **(2 points)** What are the titles of all pages linked from the page with id 42?

5. **(4 points)** What are the titles of all pages linked from pages with the title “Database”?

6. **(4 points)** Popularity of a page is defined as the number of incoming links (links leading to that page). What are the titles of the 5 most popular pages?