# CSCI585 Midterm exam

June 15th, 2017

Last Name: _____

First Name: _____

Student ID: _____

Email: _____

Signature: _____

Duration: 2 hours

CLOSED book and notes. No electronic devices.

DO YOUR OWN WORK.

If you are discovered to have cheated in any manner, you will get a 0 and be reported to SJACS.

If you continue working on the exam after time is up you will get a 0.

| Problem Set | Number of Points | Your Score |
|---|---|---|
| Q1 | 1+4=**5** | |
| Q2 | 1+1+1+2=**5** | |
| Q3 | 2+2+1=**5** | |
| Q4 | 1+2=**3** | |
| Q5 | 1+2=**3** | |
| Q6 | 2+1+1=**4** | |
| **Total** | **25** | |

Q1. (5 points total) ER Modeling

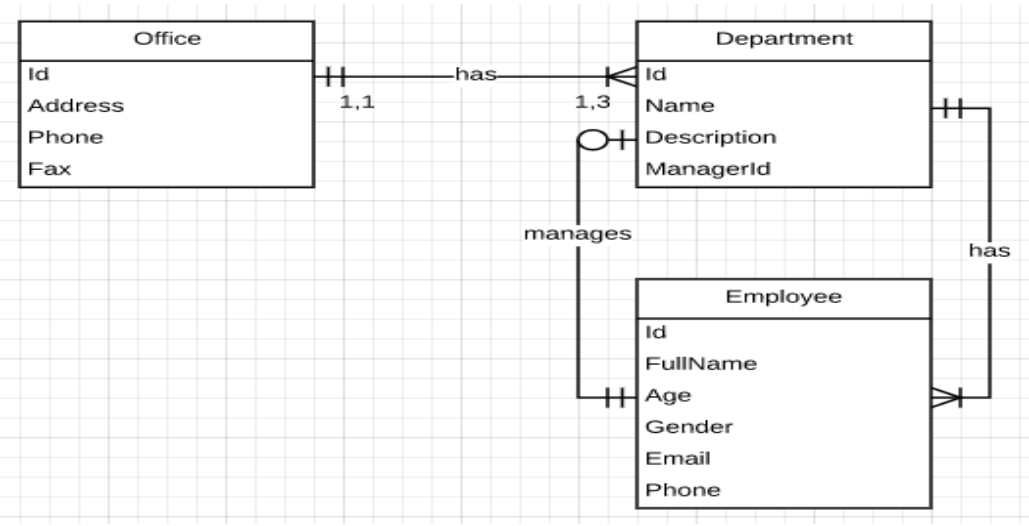
a. (1 point) What is weak entity? What is weak relationship? Give short example for each.

**Answer:** A weak entity is an entity that displays existence dependence and inherits the primary key of its parent entity. *Example: An EMPLOYEE might or might not have a DEPENDENT, but a DEPENDENT cannot exist without an EMPLOYEE.*

A weak relationship is a relationship in which the primary key of the related entity does not contain a primary key component of the parent entity. *Example: A course may have zero or many classes. A CLASS may have COURSE_ID as just a foreign key).*


b. (4 points) Draw ER Diagram based on the following description:

A company has multiple departments. Each department has multiple employees. An employee could not be in two departments at the same time. In each department, one employee is the manager that manages other employees. Each department is located at a specific office of the company. An office may have multiple departments, varying from one to three. **Answer:**



Q2. (5 points total) SQL

a. (1 point) What's the difference between inner and outer joins? Explain using example.

**Answer:** Inner join is a join operation in which only rows that meet selected criterion are selected. In outer join, all unmatched pairs are retained. Unmatched values in the related table are left null. Example: if joining product and vendor, outer join will also include products that aren't sold by vendors and vendors that aren't selling any products.

b. (1 point) Consider the following two query results:

SELECT count(*) AS total FROM books;

| Total |
|---|
| 100 |

SELECT count(*) AS author1_total FROM books WHERE authorId = '1';

| author1_total |
|---|
| 15 |

Given the above query results, what will be the result of the query below? Circle below.

```
SELECT count(*) AS author_not_1_total
FROM orders
WHERE authorId <> '1'
```

A) 50 B) 85 C) 15 D) Insufficient information

**Answer:** It will not necessarily be 85, because we do not know if there are any authorId's that are NULL. If there are 2 NULL authorId's, then the answer would be 83. Hence the answer is (d).

c. (1 point) Which relational operation does the following SQL query implement?

```
SELECT name
FROM driver
WHERE vehicle IN (SELECT vehicle FROM vehicles)
GROUP BY name
HAVING COUNT(*) = ( SELECT COUNT (*) FROM vehicles);
```

**Answer:** Division

d. (2 points) Given the following enrollment table, write a SQL query to list number of students enrolled in each course (ClassID).

| StudentID | ClassID | Grade |
|---|---|---|
| 321 | CSCI495 | B |
| 564 | CSCI110 | C |
| 321 | CSCI585 | A |
| 564 | CSCI495 | A |
| 789 | EE101 | F |
| 321 | EE101 | B |

**Answer:**

```
SELECT ClassID, COUNT(*) AS Total
FROM enrollment
GROUP BY ClassID;
```

## Q3. (5 points total) NORMALIZATION

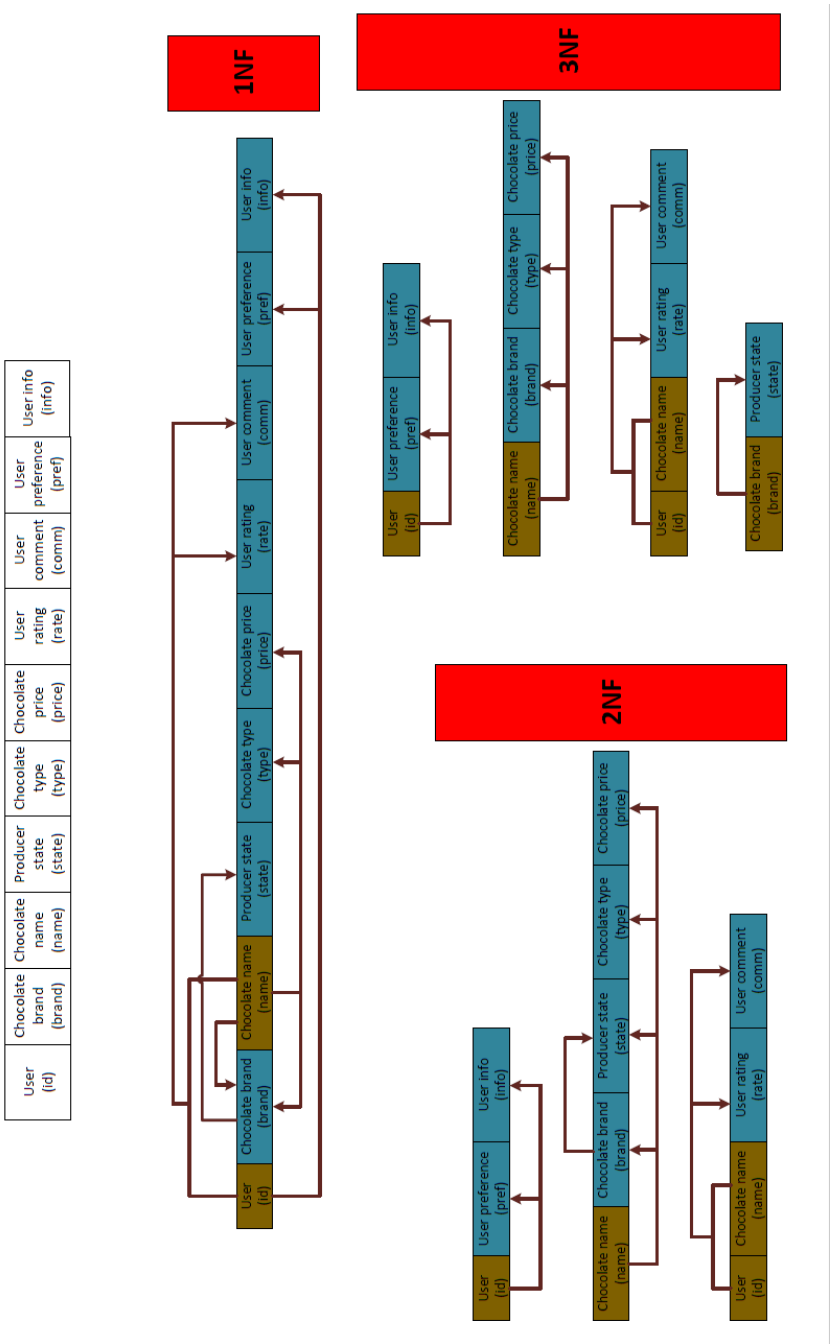
We are starting to design a new chocolate rating platform.

We have asked experts to taste different sweet products (sweets) and express their overall rating (0-5) and their professional comments. Each expert taster (ID) may try different products from different companies. We are only focused on US based companies and record the state of the headquarters of the company in our reports. Each product has a specific type, e.g. confection, candy, bar, etc. and also a unique price. Even though each taster can rate different types of products, we only record their top preference type for further reference. Moreover, basic personal information of each taster is available in our records.

Our sample data is in the following table:

| ID | Brand | Name | State | Type | Price | Rate | Comm | Pref | Info |
|---|---|---|---|---|---|---|---|---|---|
| 29001 | Hershey | Reese's | PA | Confection | 1.29 | 3 | Good peanut butter! | Bar | Michael Mast |
| 29001 | Hershey | NutRageous | PA | Bar | 1.39 | 4 | Packaging isn't great | Bar | Michael Mast |
| 29001 | Hershey | Pieces | PA | Candy | 0.99 | 2 | Worst candy ever | Bar | Michael Mast |
| 29001 | Hershey | York | PA | Confection | 0.99 | 3 | Too sweet | Bar | Michael Mast |
| 29001 | Mars | Snickers | WA | Bar | 1.45 | 3 | Not a fan of almonds | Bar | Michael Mast |
| 29001 | Mars | Twix | WA | Bar | 1.45 | 5 | Best product ever | Bar | Michael Mast |
| 29001 | Mars | M&M | WA | Candy | 1.79 | 3 | The coating is too hard | Bar | Michael Mast |
| 1202 | Hershey | York | PA | Confection | 0.99 | 3 | Just a failed cop of PBCs | Candy | Paul Bulcke |
| 1202 | Hershey | Snack Barz | PA | Bar | 0.89 | 3 | Too hard | Candy | Paul Bulcke |
| 1202 | Mars | 3Musketeers | WA | Bar | 1.69 | 2 | Best quality | Candy | Paul Bulcke |
| 1202 | Mars | Bounty | WA | Bar | 1.55 | 3 | Too much coconut | Candy | Paul Bulcke |
| 1202 | Mars | Twix | WA | Bar | 1.45 | 3 | Too much nuts | Candy | Paul Bulcke |
| 1202 | Mars | M&M | WA | Candy | 1.79 | 5 | The best taste ever | Candy | Paul Bulcke |
| 1202 | Hershey | Kisses | PA | Chips | 0.23 | 2 | Poor quality chocolate | Candy | Paul Bulcke |

a. (2 points) Identify dependencies and draw dependency diagram.
b. (2 points) Normalize our record table. Show resulting tables in 3NF.




c. (1 point) What's the purpose of DEnormalization? Give example when completely normalized table could be undesirable.

**Answer:** A process by which a table is changed from a higher-level normal form to a lower level normal form, usually to increase processing speed.

## Q4. (3 points) TRANSACTION MANAGEMENT

### a. (1 point) What are the names of the four ACID properties?

**Answer:** Atomicity, Consistency, Isolation and Durability.

### b. (2 points) Given below is the transaction log.

| TRL_ID | TRX_NUM | Table | ROW ID | Attribute | BEFORE VALUE | AFTER VALUE |
|---|---|---|---|---|---|---|
| 341 | 101 | ****Start Transaction | | | | |
| 352 | 101 | PRODUCT | 1558-QW1 | PROD_QOH | 25 | 23 |
| 363 | 101 | CUSTOMER | 10001 | CUST_BALANCE | 525.75 | 615.73 |
| 365 | 101 | ****End Transaction | | | | |

Assuming the system crashed somewhere between TRL_ID 353 – 364 (i.e, the last two rows are not in the logs), what should be the values of the two fields that were updated in that transaction after recovery?

A (25; 525.75)

B (23; 525.75)

C (23; 615.73)

D (25; 525.75)

Briefly explain your answer.

**Answer:** D

The transaction hadn't committed before the database crashed. After recovery, none of its changes should be presented. Therefore, the answer is (25; 525.75)

Q5. (3 points) OPTIMIZATION

a. (1 point) Given below are two queries that perform the same function. Which one do you think would be more efficient, and why?

```
SELECT id, name
FROM viterbi_Students
WHERE branch = 'Computer Science' AND courseTaken = 'CS 585';
```

```
SELECT id, name
FROM viterbi_Students
WHERE courseTaken = 'CS 585' AND branch = 'Computer Science';
```

**Answer:** The second one is more optimized as in AND, we should write the condition that is more likely to be false first. Clearly, the number of students taking CS 585 would be much lesser than the number of students who are enrolled in Computer Science.

b. (2 points)
In a smartphone store, where roughly 20000 smartphones are added to the inventory every week, what recommendation would you give the designer about the use of derived attributes? Write an improved query based on your assumption.

```
SELECT model.modelCompany, AVG(model.modelPrice *
company.ratingCompany)
FROM model INNER JOIN company ON model.modelCompany = company.name
WHERE model.modelPrice > 400
GROUP BY model.modelCompany;
```

**Answer:**
Consider having an attribute in model which stores model.modelPrice * company.ratingCompany.

```
SELECT modelCompany, AVG(priceBase)
FROM model
WHERE modelPrice > 400
GROUP BY modelCompany;
```

Q6. (4 points) DISTRIBUTED DATABASES

Every year a large venture capital company needs to invest many projects. They use a table PROJECT to keep track of each project. However, at the end of each year, their financial department needs to check how much money they have invested. Recently, they are considering to change to a distributed database to store the data.

PROJECT

| Project_id | PName | Budget | Location | Manager |
|---|---|---|---|---|

a. (2 points) Which data fragmentation technique should they use to meet the requirement of the financial department? Show new design.

**Answer:** Vertical fragmentation.

| Project_id | Budget |
|---|---|

| Project_id | PName | Location | Manager |
|---|---|---|---|

b. (1 point) If budget department were to have multiple branches in different states to manage their local projects, which data fragmentation technique should they use?

**Answer:** Horizontal fragmentation. Each sub-table stores its state's project records (rows).

c. (1 point) Which protocol is used to control the distributed concurrency (by DDBMS)?

**Answer:** **DO-UNDO-REDO protocol**: Roll transactions back and forward with the help of the system's transaction log entries.