CMSC498O
Fall 2014

Introduction to Data Science I
MIDTERM
Closed Book

Deshpande

- Total points: 80. Weight: 15% of the course grade.
- Show your reasoning. Write partial solutions. You will get a fair amount of the credit if I think you know the concepts.
- Unless otherwise specified, a *Yes/No* answer is *not sufficient* for any question. No points will be given without accompanying explanation.

Your Name:

Miscellaneous Questions 1 (9 questions - 2 pt each)

1. What is "volunteer bias" in sampling? You can use an example.

2. What are "protocol buffers"? What are they used for?

3. List three different data models.

4. What is the difference between “deduplication” and “record linkage” in the context of entity resolution?

5. Briefly explain the “local-as-view” approach in data integration.

6. Explain the rule-based approach to *relation extraction* in Information Extraction with an example.

7. Briefly explain the notion of “regularization” in statistical modeling.

8. List and briefly explain one classification technique.

9. Consider the relations: $R(A, B)$ and $S(B, C)$, and the SQL query:

   ```
   select R.a, count(*) from R natural join S group by R.a;
   ```

   Briefly explain the result of this query in words. Why might you want to use *left outer natural join* instead of *natural join* here ? Assume A is a primary key for R.

Miscellaneous Questions 2 (9 questions - 3 pt each)

1. What is the wrong with the following statistical analysis?

2. Show how to compute the p-value for the following data. You can write the formula and leave it at that.

3. Organizational data with abbreviations. Give some approaches to disambiguate. Which of the following is likely to give best results. Explain your reasoning.

4. On the following tables, what are the results of the queries listed below?

R

A	B	C
'a1'	10	10
'a1'	20	20
'a2'	30	30
'a2'	0	NULL

S

C	D
30	'd1'
NULLL	'd2'

- `select avg(B) from R group by A`:

- `select * from R where C != 10`:

- `select * from R, S where R.C = S.C or R.C is null`: The result contains three tuples.

5. Fill in the pseudocode for a naive implementation of the aggregation operation in the following query using Hashing.

   ```
   select R.A, sum(R.B) from R group by R.A
   ```

   ```
   ---------------------------------------------------------------------------
   HashMap h = new HashMap();
   for each tuple r in R:









   // print out the results









   ---------------------------------------------------------------------------
   ```

6. What does 'I' stand for in ACID properties? Briefly describe one mechanism for ensuring 'I'.

7. List and briefly describe three single-source data quality problems.

8. Consider the following schema:
   create table r (a integer primary key, c integer);
   create table s (b integer primary key, a integer references r);
   create table t (c integer primary key, b integer references s);
   alter table r add constraint rreft foreign key (c) references t(c);

   - Why can't I add the foreign key reference directly in the "create table" statement for table "r" ?

   - Explain why the statement "drop table r" would be rejected.

   - Is there any way I can delete all the tables ? Explain in words.

9. SQL 1: The following two queries are not equivalent (they don't always produce identical results) because of NULLs. Identify and explain the problem. Schemas are: $R(a,b,d), S(c,d)$. Assume a is the primary key for R.

Query I	**Query II**
select a **from** R **where** R.b = (**select count**(S.c) **from** S **where** R.d = S.d)	**select** a **from** R, S **where** R.d = S.d **group by** R.a **having** R.b = **count**(S.c);