FINAL EXAM, 839 SPRING 2019

Write down your name here (last name first) AND student ID number:

Please wait until you are told to start working on the exam.

75 minutes, closed books and notes, no electronics, one page cheat sheet allowed.

If a question is ambiguous, please state your assumptions and then answer based on those assumptions. Please keep the assumptions reasonable.

You can only use the space provided in this exam. Adding more space is not necessary and may result in points being deducted.

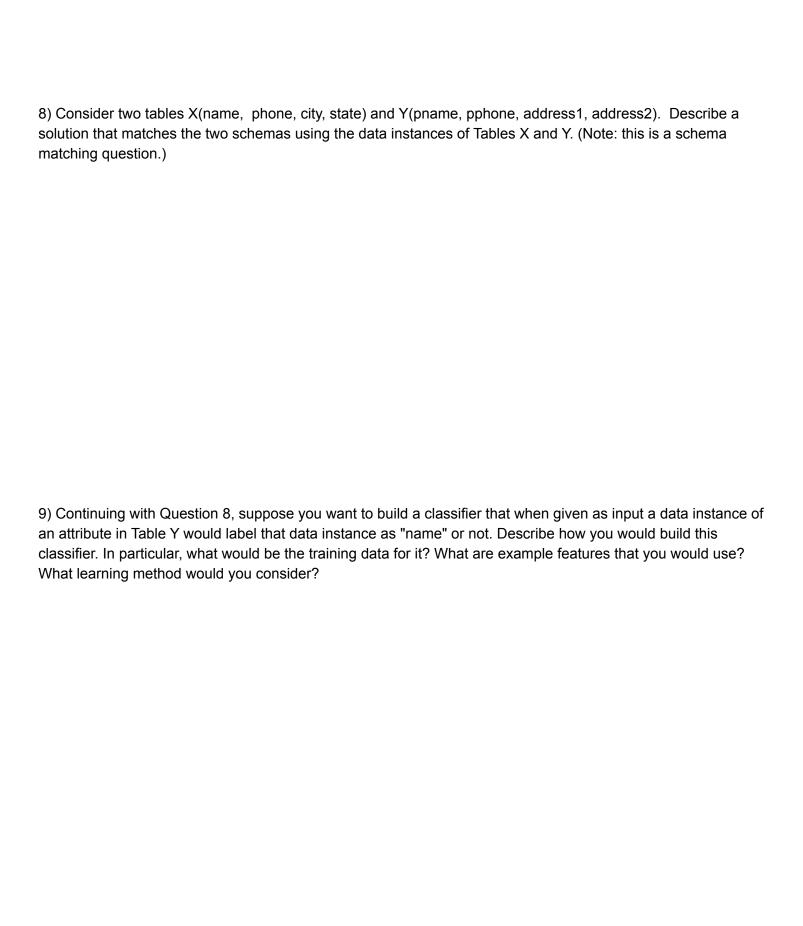
Some questions may be harder than others. You may want to handle easier questions first, delaying the more difficult questions to later.

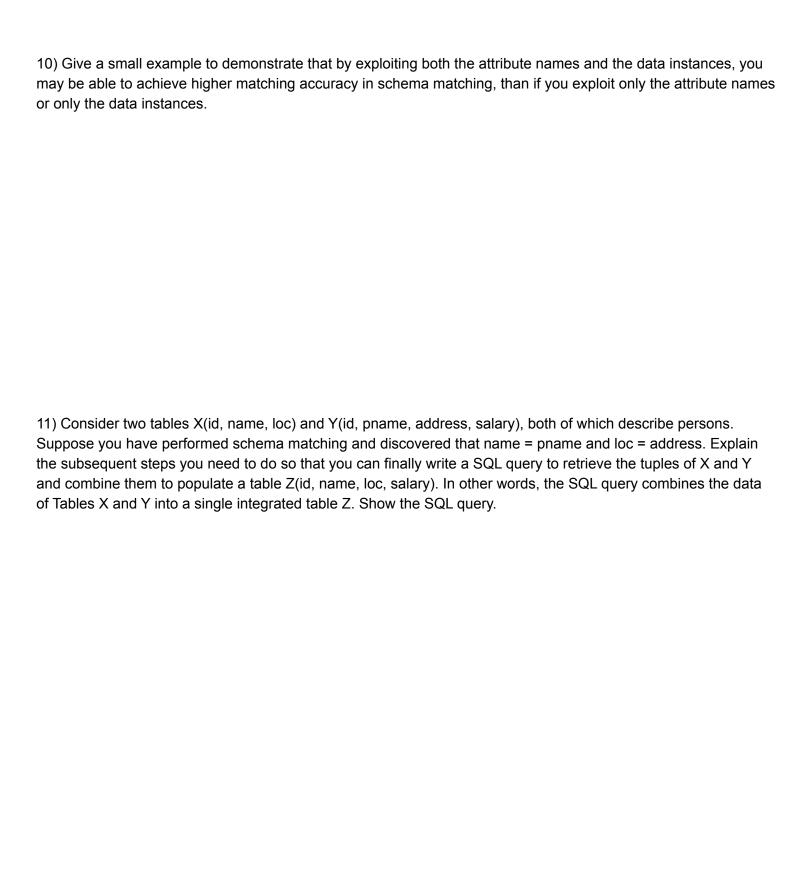
15 Questions, 5 Points Each, for a Total of 75 Points.

1) Suppose you want to match tuples from two tables A and B, where each tuple describes a person. Suppose the schemas of these tables are the same: A(name,city,state,age) and B(name,city,state,age). Define the purpose of the blocking step, and give at least two examples of blocking that you can do on these two tables.
2) Suppose you want to do blocking on "city", that is, if two tuples do not agree on "city", then the pair consisting of these tuples does not survive the blocking step. What happens if you have a lot of missing values in "city"? What would you do?

3) Continuing with Question 2, besides missing values, are there any other issues you should worry about before trying to do blocking on "city"? Explain your answer.
4) Suppose you want to go ahead to do blocking on "city", as described in Question 2. Describe an algorithm to do this blocking step fast.
5) Suppose you have a table where each tuple describes a book. Suppose also that the attribute "publisher" has a lot of values such as "Springer", "Springer-Verlag", "Springer Publisher", etc., in short, many values referring to the same publisher but using different variations. You want to clean up this attribute, that is, normalize all the values so that all strings referring to the same publisher will be normalized to a single canonical string (e.g., "Springer-Verlag"). Briefly discuss a solution to do this.

6) Suppose you want to perform entity matching us example where it is more appropriate to use than us scores.		
7) Why schema matching is difficult?		





14) Briefly describe a star schema (in OLAP)
15) Explain and contrast OLTP and OLAP queries.
13) Explain and Contrast Offic and Offic quenes.