CS 3200: Homework 1
Prof. Rachlin

WHAT TO SUBMIT: Submit a single PDF containing your answers to part A and a single text file containing your program or pseudocode.

PART A (80 POINTS): Answer each of the following questions (10 points each)

1. Discuss the differences between DDL and DML. What operations would you typically expect to be available in each language?

2. Describe the difference between data security and data integrity.

3. What advantages do databases provide over early file-based approaches? Give at least three examples.

4. What is concurrency control and why does a DBMS need a concurrency control facility?

5. What is a transaction? Give an example of a transaction.

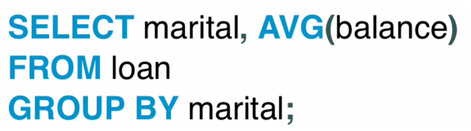
6. What is meant by the term 'client-server architecture' and what are the advantages of this approach? Compare the client-server architecture with two other architectures.

7. In the 2016 InfoWorld article "What eBay looks like under the hood" what problem did eBay have with their product catalog? How is this a database design problem? Compare eBay's approach to the way in which Amazon allows third-party vendors to sell used books.

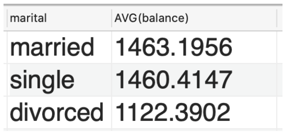
8. The two assigned articles from the February issue of The Scientist both mention the building of Databases. How do these databases facilitate the scientific research discussed in each article? Is the rational for building a database in each case similar? Justify your answer.

PART B (20 POINTS): In big data environments, analysts are constantly writing *ad hoc* database queries to obtain general table statistics for purposes of validation, trend analysis, and to simply gain a deeper understanding of their data. Suppose we had a .CSV text file listing bank customer loan applications. Each line includes data about the customer: age, marital status, education, account balance, and so on. The last column indicates whether the customer was approved for the loan.

The bank manager would like to know if customers with different marital status have, on average, significantly different account balances. In a programming language of your choice (or using pseudocode), write or describe a method that reads through the CSV line-by-line, and outputs the average account balance for customers in each marital status category. The equivalent SQL query, and expected output, is provided below. (Your program will probably require more than three lines of code. Using Python pandas or R dataframes is not allowed – nice try!)



```
SELECT marital, AVG(balance)
FROM loan
GROUP BY marital;
```



marital	AVG(balance)
married	1463.1956
single	1460.4147
divorced	1122.3902

What difficulties are overcome using SQL and a Database Management System?