1. **(Questions 1-3 pertain to the class lecture "Exploring the Relational Data Model of CSV")**
   What is the estimated population of La Paz county in the state of Arizona? (Choose the best answer.)

   - ○ 25000
   - ◉ 20000
   - ○ 15000
   - ○ 10000

2. What county in the state of Wyoming has the smallest estimated population?
   - ○ Unita
   - ○ Sweetwater
   - ○ Platte
   - ◉ Niobrara

3. In the class we created a filter for all of the counties in California with a population greater than 1000000. However, included in the results is the entire state of California. This anomalous value might skew our analysis if, for example, we wanted to compute the average population of these results. What additional filter(s) might work to resolve this problem?
   - ☑ Add a filter to detect and remove results which do not include the word "California" in column G.
   - ☑ Add a filter where the value in column E is greater than zero (0).
   - ☐ None of the above
   - ☐ Add a filter to detect and remove results which do not include the word "county" in column G.
   - ☐ Add a filter which finds all counties with population greater than 1,000,000 AND less than 10,000,000 for column H (CENSUS2010POP).

4. **(Questions 4 and 5 pertain to "Exploring Sensor Data")**
   How often (in seconds) do the R5 measurements occur?

   - ○ 30
   - ○ 50
   - ○ 40
   - ◉ 60

5. What is the field for rain accumulation?
   - ○ Dn
   - ○ Sm
   - ○ Dx

- (●) Rc

6. **(Questions 6 and 7 pertain to "Exploring the Array Data Model of an Image")**
   What is the (Red, Green, Blue) pixel value for location 500, 2000?
   - ( ) (100, 123, 149)
   - ( ) (134, 145, 46)
   - (●) (163, 118, 79)
   - ( ) (50, 156, 182)

7. Is this value likely to be land or ocean?
   - ( ) Ocean
   - (●) Land

8. **(Questions 8 and 9 pertain to "Exploring the Semistructured Data Model of JSON")**
   Given a tweet, what path would you most likely enter to obtain a count of the number of followers for a user?
   - (●) user/followers_count
   - ( ) user/statuses_count
   - ( ) user/listed_count
   - ( ) None of the above

9. Which of the following fields are nested within the 'entities' field (select all that apply)?
   - [x] user_mentions
   - [ ] views
   - [ ] events
   - [ ] tweets
   - [x] symbols
   - [x] urls

10. What is a possible pitfall of utilizing Excel as a way to manipulate small databases?
    - ( ) Excel is a user program and thus cannot run on a server.
    - (●) Excel does not enforce many principles of relational data models.
    - ( ) Excel does not allow algorithms for data manipulation.
11. What does the term "atomic" mean in the context of relational databases?
    - ( ) A column or row of data. Depends on the context.

- Fixed schema of a particular database.
- One unit of information that cannot be decomposed.
- A tuple that cannot be reduced.

12. What is the Pareto-Optimality problem?
    - Find the optimal path that requires going through specific nodes given by the user.
    - Find the best possible path given two or more optimization criteria where neither constraint can be fully optimized simultaneously.
    - Find the shortest path from source node to target node.
13. What constitutes a community within a graph?
    - A neighborhood defined by an integer constant K around a specific node. All K+1 nodes belong in another community.
    - High density of nodes at a certain location.
    - Many anomalous neighborhoods within the same vicinity.
    - A dense amount of edge connections between nodes in a community and a few connections across communities.
14. Why are trees useful for semi-structured data such as XML and JSON?
    - It is not always the case that XML and JSON can be represented as trees.
    - Trees take advantage of the parent-child relationship of the data for easy navigation.
    - They are only useful for XML data as tree-like structure is apparent with tags. While JSON does not contain a tree-like structure as it contains arrays.
    - Computers can easily visualize the data with a tree structure.
15. What is the general purpose of modeling data as vectors?
    - Enables image searching.
    - Results can be ordered by similarity using vector projection.
    - Enables weighting of the query.
    - The ability to normalize vectors allowing probability distributions.
16. For the following questions 7, 8, and 9, suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number.
    Suppose we collect data month by month. Each month, we would have a batch of data containing the fields listed above. At the end of the year, we would remove redundancies in our data by removing any records with duplicate account numbers. What type of operation do we use in this scenario?
    - Not an Operation
    - Union
    - Sub setting
    - Join

17. From the information given in question 7, what are the constraints, if any, which we have placed on the Account Number field for the end of year collection?
    - ○ Account should have at most n digits.
    - ○ There are no constraints.
    - ◉ Account Number should be unique.
    - ○ If we had n duplicate Account Numbers then we will remove n-1 duplicate fields.
18. Suppose 100 people signup for our system and of the 100 people, 60 of them did not input an address. The system lists the values as NULL for these empty entries in the address field. Would this situation still have structure for our data?
    - ○ No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.
    - ◉ Yes the data has structure because we have placed a structural constraint on the data, thus the data will always have the originally defined structure.

19. What is true between data modeling and the formatting of the data?
    - ○ There is always one specific schema for storing model data that is the best and preferred method for the specific data representation.
    - ○ The data format has no relation with data modeling. The way data is stored has no correlation to how data is displayed.
    - ◉ The data does not necessarily need to be formatted in a way that represents the data model. Just so long as it can be extrapolated.
    - ○ There is a one to one correspondence between formatting data and data modeling. For every model of data, there is only one way to store the data.

20. What is steering?
    - ◉ Utilizing real time data to compute and change the state of an application continuously.
    - ○ Using static data stored from a real time source in order to process and guide the application.
    - ○ Using sensors to manipulate the system, such as a smart car being able to drive by itself using sensors to detect road hazards.
    - ○ Calculating results using real time data otherwise known as streaming data.

21. Of the following, what best describes the properties of working with streaming data?
    - ☑ Small time windows for working with data.
    - ☑ Does not ping the source interactively for a response upon receiving the data.
    - ☐ Data is always utilized for steering the application.
    - ☑ Independent computations that do not rely on previous or future data.

- [ ] Always unbounded in sequence, in other words, data is not guaranteed to be in order.
- [x] Data manipulation is near real time.

22. What is a characteristic of streaming data?

- ( ) Data is finite in size and size determines the time and space of processing the data.
- ( ) The data is finite and requires only finite time and space to process the data.
- (•) Data is unbounded in size but requires only finite time and space to process it.
- ( ) The data is unbounded in size and the size determines the time and space of processing the data.

23. What type of algorithm is required for analyzing streaming data?

- ( ) Accurate and Memory Efficient
- ( ) Fast and Complex
- ( ) Accurate and Consistent
- (•) Fast and Simple

24. What is lambda architecture?

- (•) A method to process streaming data by utilizing batch processing and real time processing.
- ( ) A specific hardware architecture for a server made specifically for processing real time data.
- ( ) A specific method for processing streaming data using special real time processes.

25. Of the following, which best represents the challenge regarding the size and frequency of data?

- (•) The size and frequency of the data may be sporadic.
- ( ) There may not be data to produce the notion of size and frequency.
- ( ) The size and frequency of the data may be too small.

26. What is the difference between data lakes and data warehouses?

- ( ) Data lakes utilize hierarchical systems while data warehouses use object storage.
- (•) Data lakes house raw data while data warehouses contain pre-formatted data.
- ( ) Data lakes house larger volumes of data than data warehouses.

27. What is schema-on-read?

- (•) Data is stored as raw data until it is read by an application where the application assigns structure.
- ( ) The process where formatted data is given structure when read.

- Another name for data lakes.
- The process where data is pre-formatted prior to being read but the schema is loaded on read.