
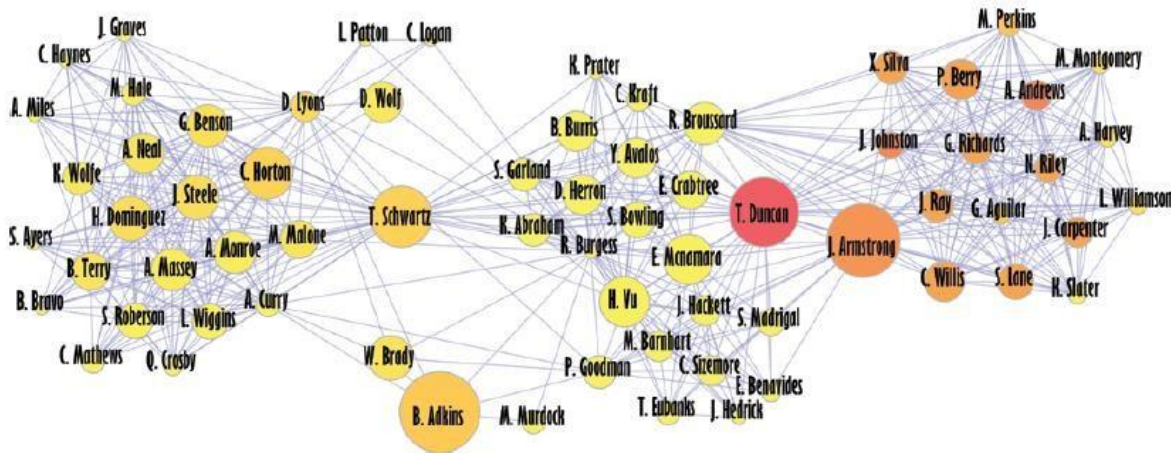


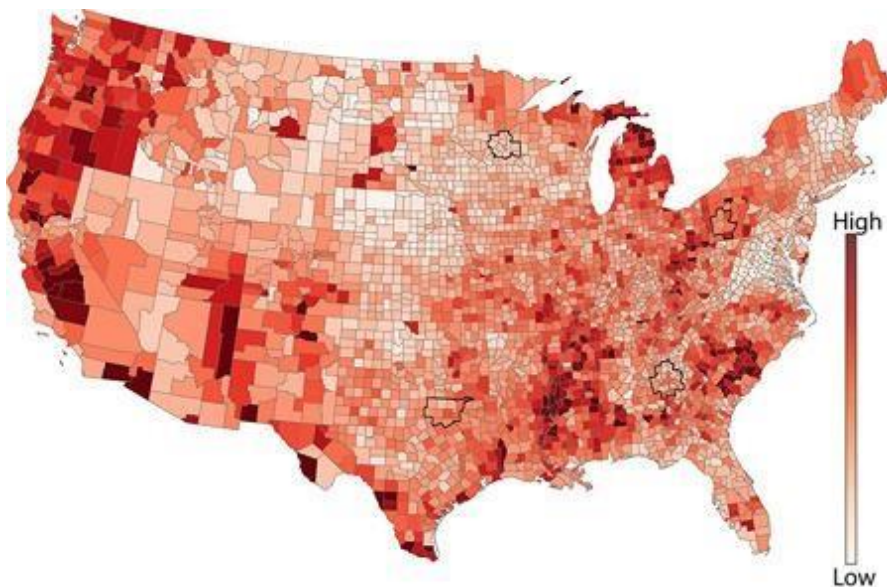
CSCI 585 QUESTION BANK- Midterm 2

Q1. Name the type of depiction-

1. The following depicts the links between people on a Social Network.
What is this type of depiction called? 





2. The following depicts spatial data covering an entire region. What is this type of depiction called? 



Q2. Consider the following table-

EMPLOYEE ID	EMPLOYEE NAME	DEPARTMENT	PHONE NUMBER	TYPE
1	Alia	IT	25321425	Home
1	Alia	IT	58620568	Mobile
2	Ken	Finance	56300023	Work
3	John	Sales	55853001	Work
3	John	Sales	58620432	Mobile

Complete the following json using the above data-

```
{  
  {  
    "id" : 1,  
    "name" : "Alia",  
      
  },  
  {  
    "id" : 2,  
      
  }  
}
```

}



Q3. I want to form a recommendation engine and am contemplating on which database to use. Suggest a type of NoSQL database that I should and the reason for the same.



Q4. How would you describe Data Science in your own words, in a sentence or two? You need not give the definition of Data Science. Just explain your understanding about Data Science.



Q5. Data Analysis is a huge process that starts with Data Collection. Explain the lifecycle of Data that you understand.



(HINT: Explain the lifecycle of data from collection to cleaning to storage to analysis and the visualization and reposting)

Q6. Why do you think Map Reduce is fast? Also what does YARN specifically do?

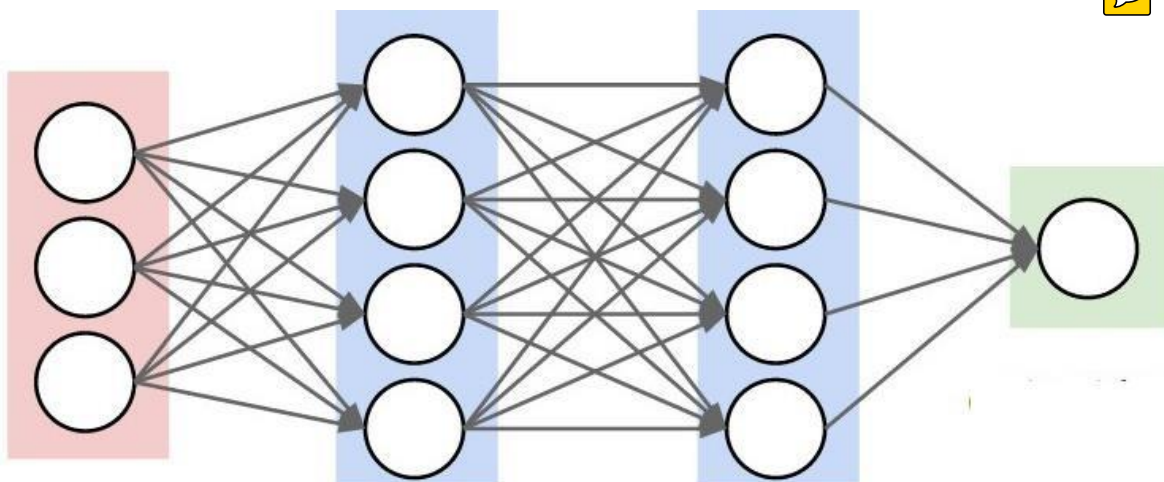


(HINT: MapReduce used SIMD implementation)

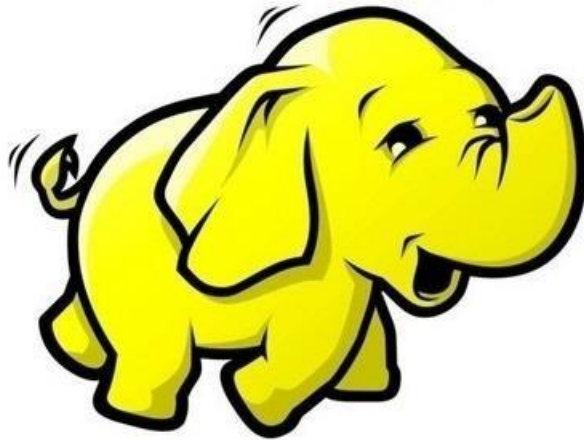
Q7. A lot of analysis work is done using R programming in major companies only because of the high level datatypes in R. Name two such high level datatype in R.



Q8. The following represents a very famous Data Mining algorithm. Identify the same and explain in your own words how does it learn?



Q9. The following toy has the same name of a new technology that increased the data processing speed to a great extent. Name the technology and explain why is it so fast?



Q10. The following is a screenshot of an algorithm being run in WEKA. Identify the algorithm run.



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose :

Cluster mode

- ☒ Use training set
- ☐ Supplied test set Set...
- ☐ Percentage split % 66
- ☐ Classes to clusters evaluation (Num) Purchase
- ☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

16:41:23 - SimpleKMeans

Clusterer output

Missing values globally replaced with mean/mode

Final cluster centroids:



Attribute	Full Data (100.0)	Cluster# 0 (26.0)	1 (27.0)	2 (5.0)	3 (14.0)	4 (28.0)
Dealership	0.6	0.9615	0.6667	1	0.8571	0
Showroom	0.72	0.6923	0.6667	0	0.5714	1
ComputerSearch	0.43	0.6538	0	1	0.8571	0.3214
M5	0.53	0.4615	0.963	1	0.7143	0
3Series	0.55	0.3846	0.4444	0.8	0.0714	1
Z4	0.45	0.5385	0	0.8	0.5714	0.6786
Financing	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Q11. What role do minimum bounding rectangles (MBR's) play in spatial query processing? (How are they used/helpful)



 Q12. What is the file format that is used to read a table in Weka? Name the file format and provide a small example of the same. 

 Q13. There was a fraud that occurred with the customers of XYZ bank. The bank has hired an agency to detect the frauds. The agency uses a binary classifier to analyze card transactions. Name two algorithms that can be used for this purpose. 

 Q14. Name the popular capable module that can be used for Tensor Flow.  

Q15. There is a Machine Learning algorithm that divides the data points into two sets by a vector plane. State the algorithm and also give an application of the same. 