Exercise 15 - Midterm awareness exercise

Instead of recapitulating todays events and clustering things back and forth, we will instead dedicate this exercise the upcoming midterm and reflect on the topics covered.

1)Big data and the three Vs (or was it four Vs)? What are they? Examples and what it means to big data.

2)Streaming, brute force, distributed computation, the cloud and wasn't there a streaming algorithm we talked about?

3)Right, data needs to be preprocessed. Binning, missing data, and how to deal with it.

4)SQL our friendly database: How to make one? How to enter data? How to retrieve it, and a couple of extras like: SUM, COUNT, MIN, MAX, AVERAGE….

5) Distance measures! What are they? Are there different ones? Euclidean distance isn't everything, what about Jaccard, simple matching coefficients, cosine?

6) We fitted data using this method called Least Square Error or something. How did it work again? Also, how can I tell that I am probably using the right function to fit?

7) Decision trees! What are they, how do they work? And there was something called GINI index and ENTROPY, how do you use them to make trees again?

8) Association Rule Mining is similar to the above. What is confidence, what is support, why is that important? How do you use it, why is this much easier than one would think?

9) Clustering! What does it do? Why are the different approaches? MIN, MAX, CURE are just buzzwords or do they mean something specific?