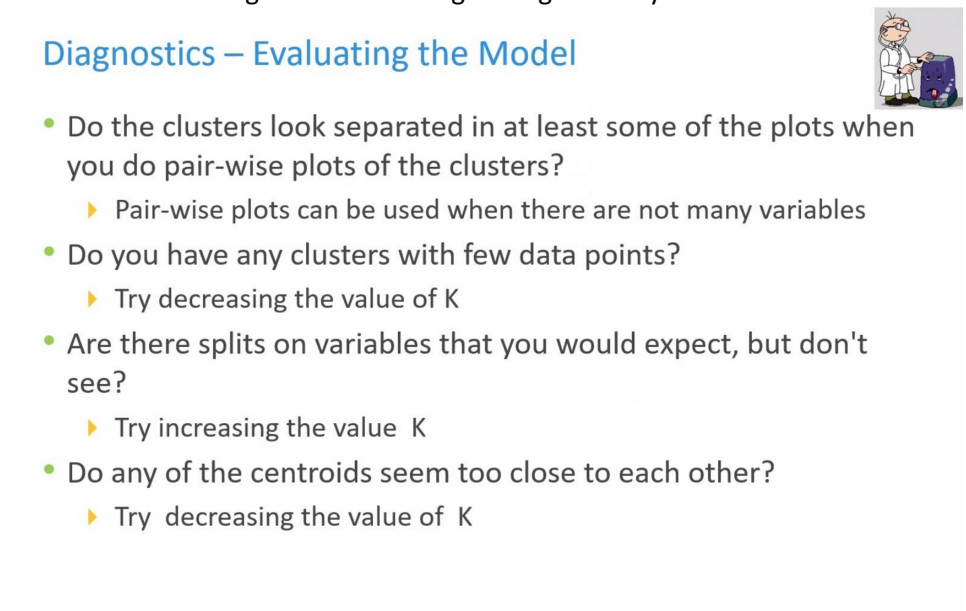
# Lec 10

Thursday, 7 November 2024 19:33

1. ELbow method for K picking in K means. WSS - sum of squared distances of records to picked centroids. Minimizing while wss changes insignificantly. Better recall from lecture

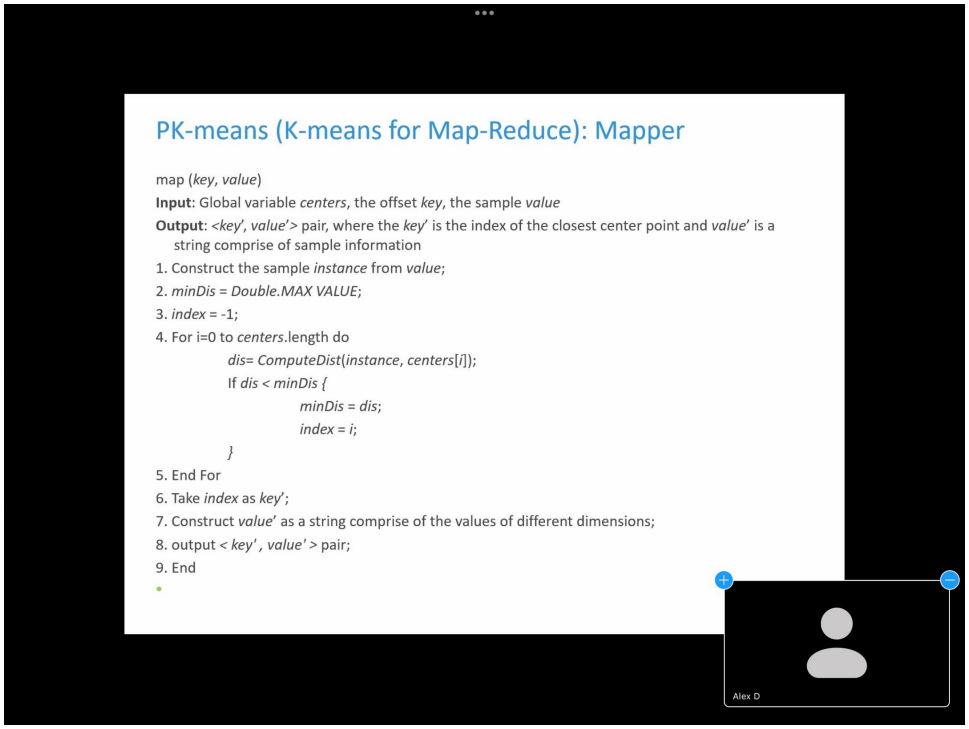
2. 


3. K-Means Clustering - Reasons to Choose (+) and Cautions (-)

| Reasons to Choose (+) | Cautions (-) |
|---|---|
| Easy to implement | Doesn't handle categorical variables |
| Easy to assign new data to existing clusters<br>Which is the nearest cluster center? | Sensitive to initialization (first guess) |
| Concise output<br>Coordinates the K cluster centers | Variables should all be measured on similar or compatible scales<br>Not scale-invariant! |
| | K (the number of clusters) must be known or decided a priori<br>Wrong guess: possibly poor results |
| | Tends to produce "round" equi-sized clusters.<br>Not always desirable |

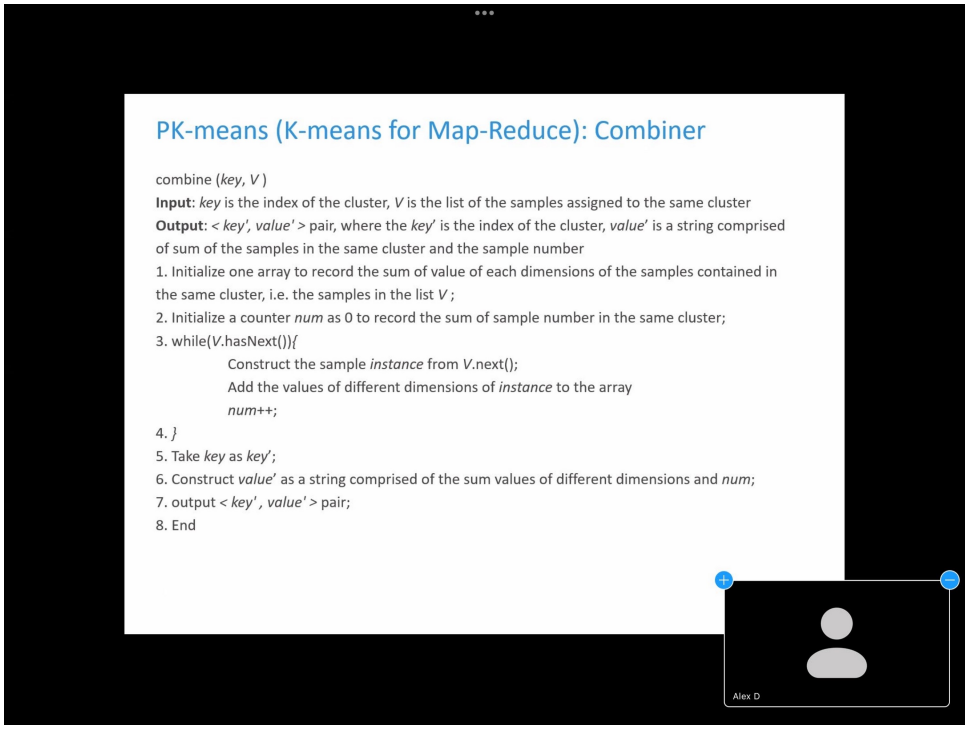
1. The map phase
2. map_of_points Map(centroid_coordinates, all_points) - returns a map of key-value pairs that represent assignment of points to closest centroids
3. new_centroid_coordinates Reduce(map_of_points) - returns new recalculated centroids based on every cluster formed

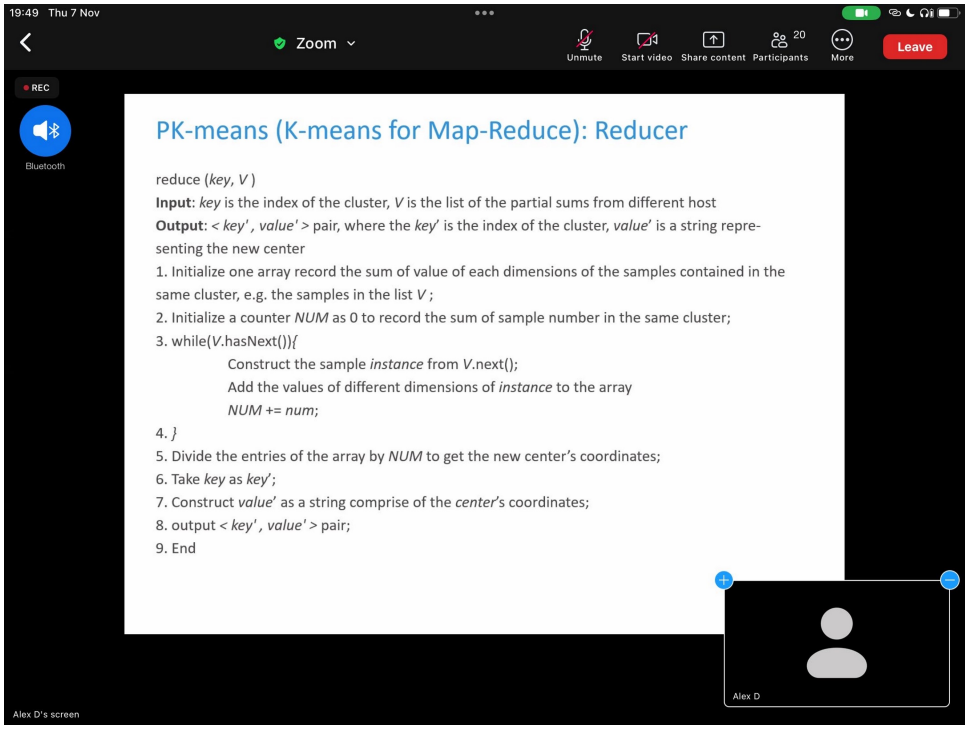
4.

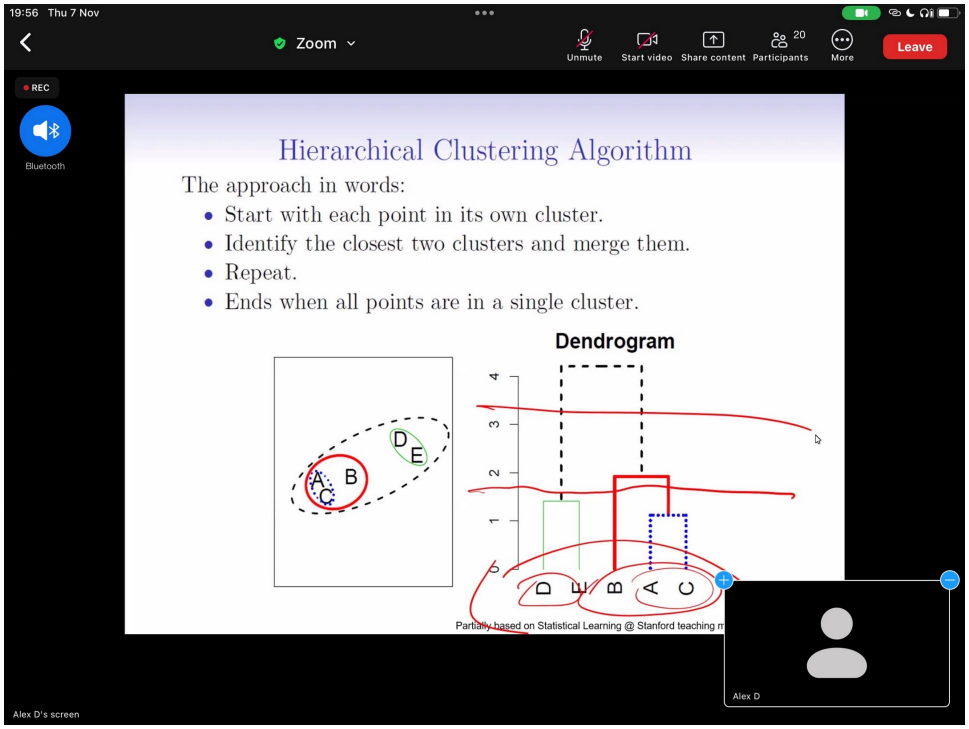



5. Combiner is sort of middleware in this pipeline that partially takes up computation from reducer

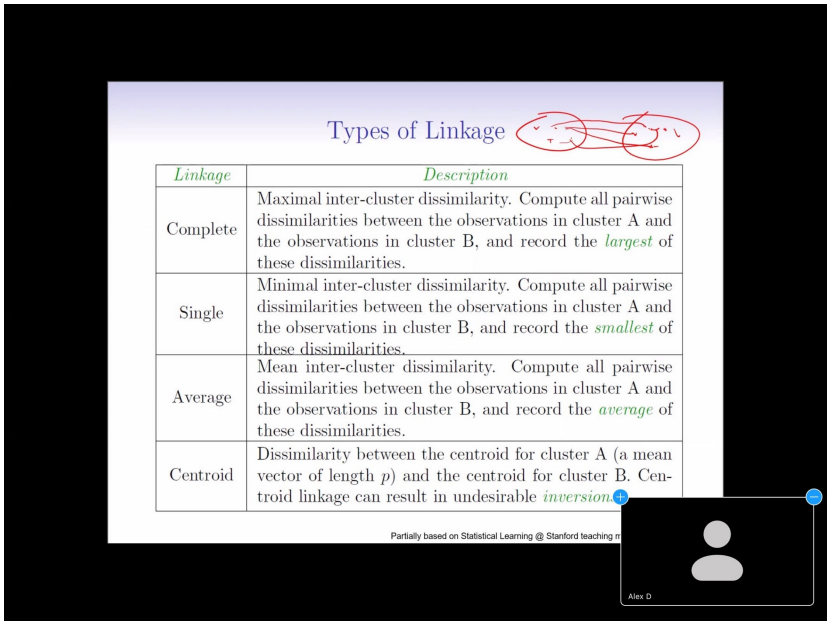
6.

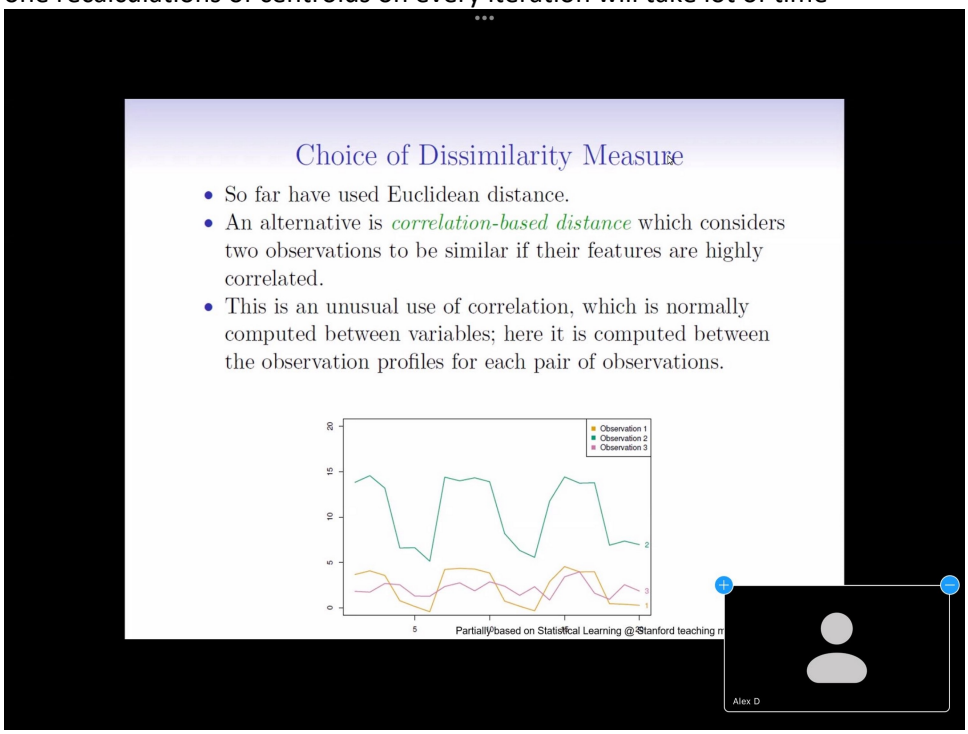


7. 

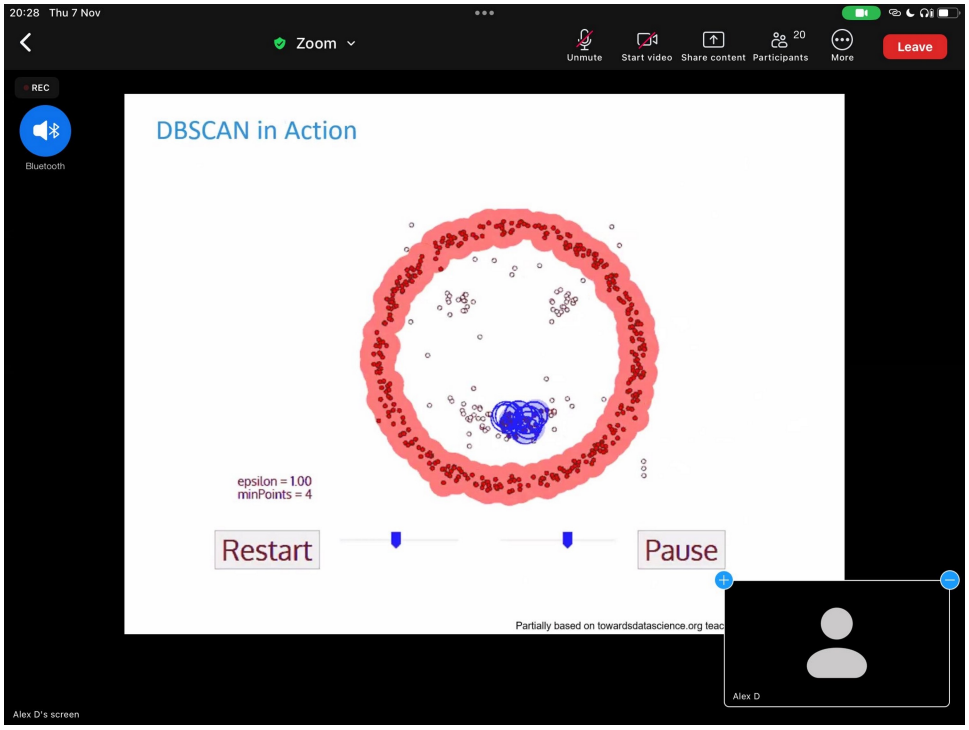

8. 

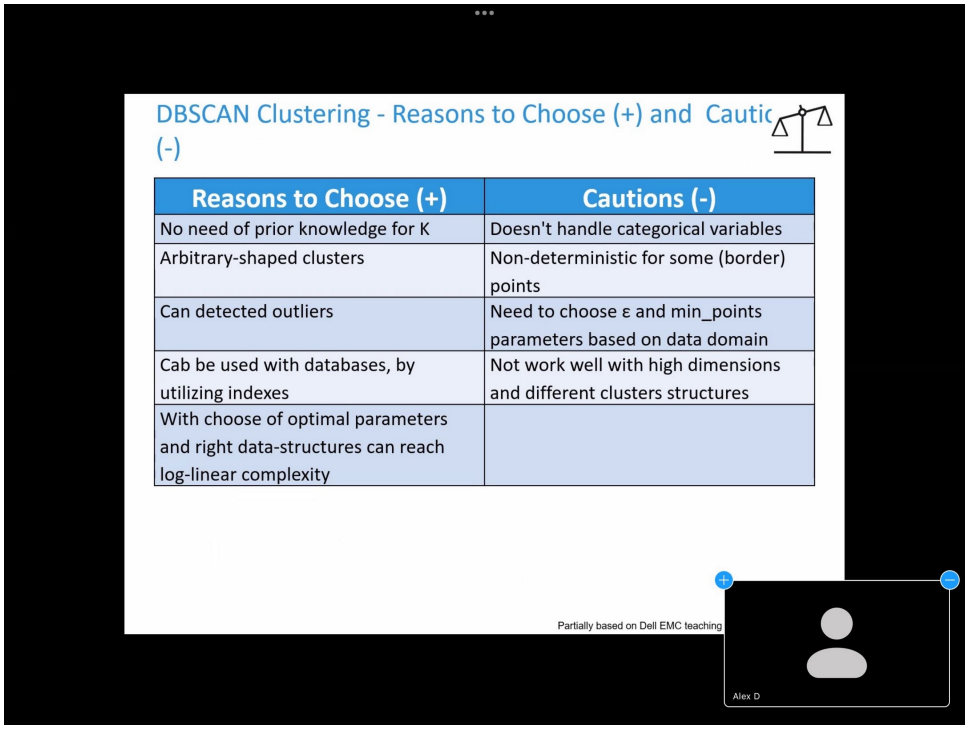

9. Hirerchial clustering is like agario lol
10. What linkage to choose to determine distance between clusters?

11. 

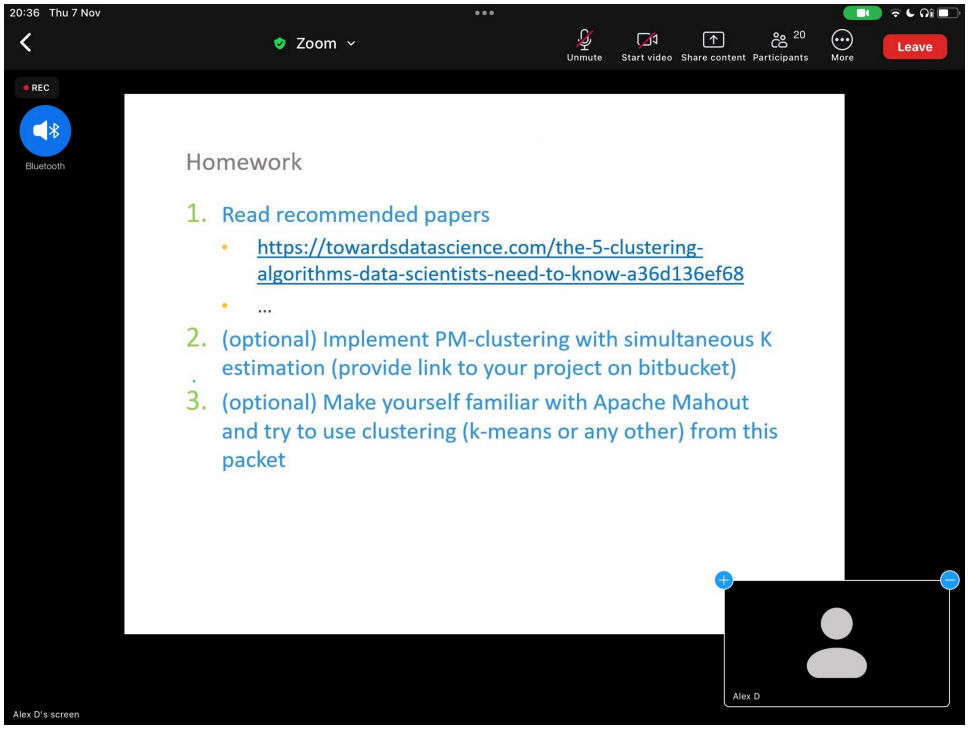
12. Better not use with lots of data. In first 3 the matrix of distances would be too big, in the last one recalculations of centroids on every iteration will take lot of time
13. 

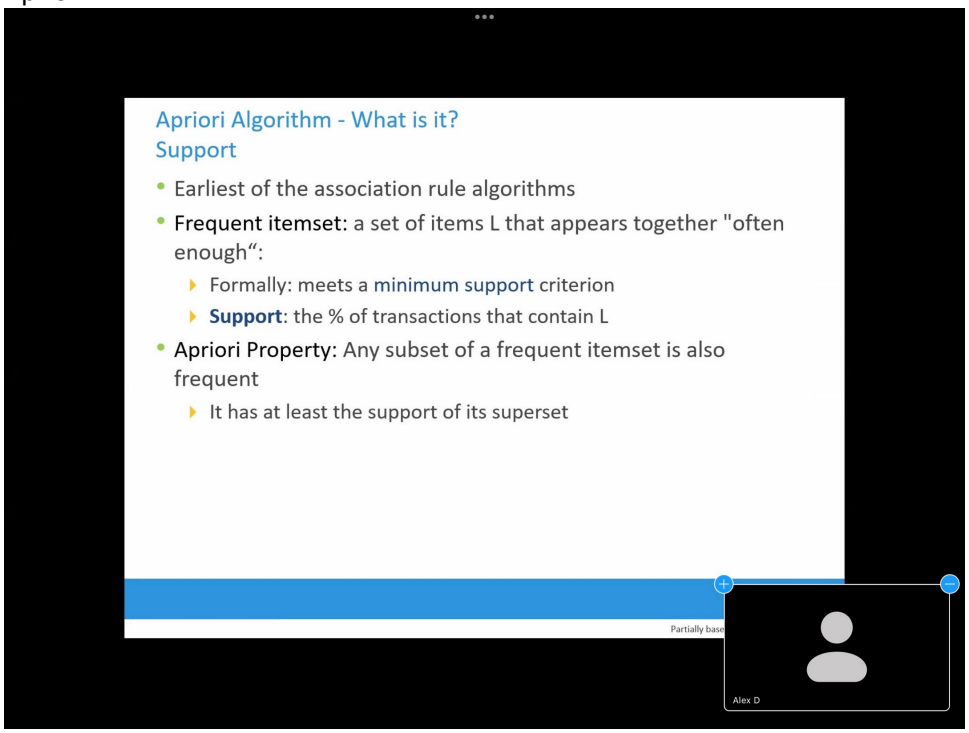
14. DBSCAN
15. Точка опорная если её количество соседей больше чем заявленное в эпсилон окрестности
16. Связываем её опроных соседей к тому же кластеру что и она
17. Затем добавляем все точки которые захватываются опорными, но сами опорными не являются(на них формирование кластера останавливается)
18. И так волной для граничных точек, рассматривая их как опорные пока такие точки не закончатся
19. Закончится всё на граничных точках - точках которые лежат в эпсилон окрестности опорной, но сами опорной не являются
20. Шумовая точка - не граничная и не опорная

21. 

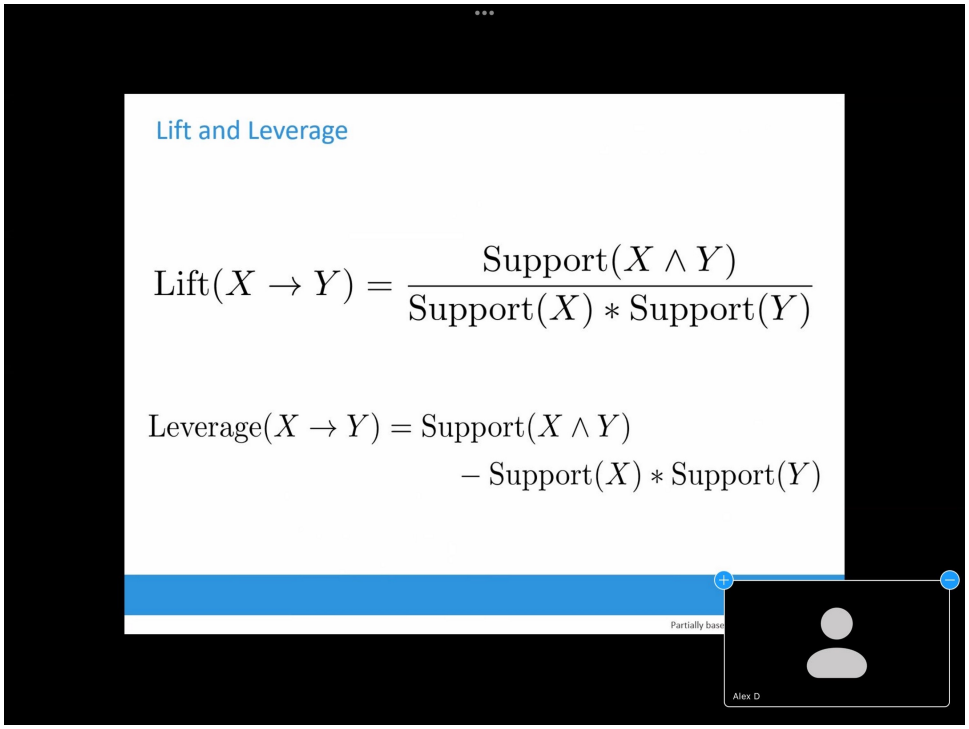

22. 


23. Additional readings

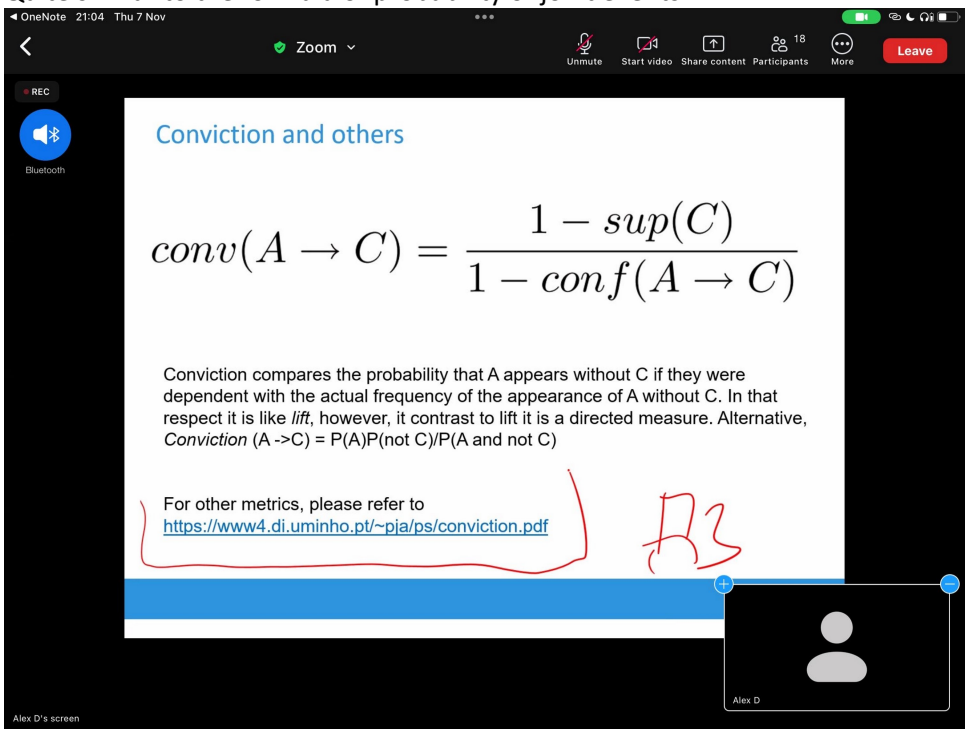
24. 

25. Assossiatin algos
26. Apriori
27. 

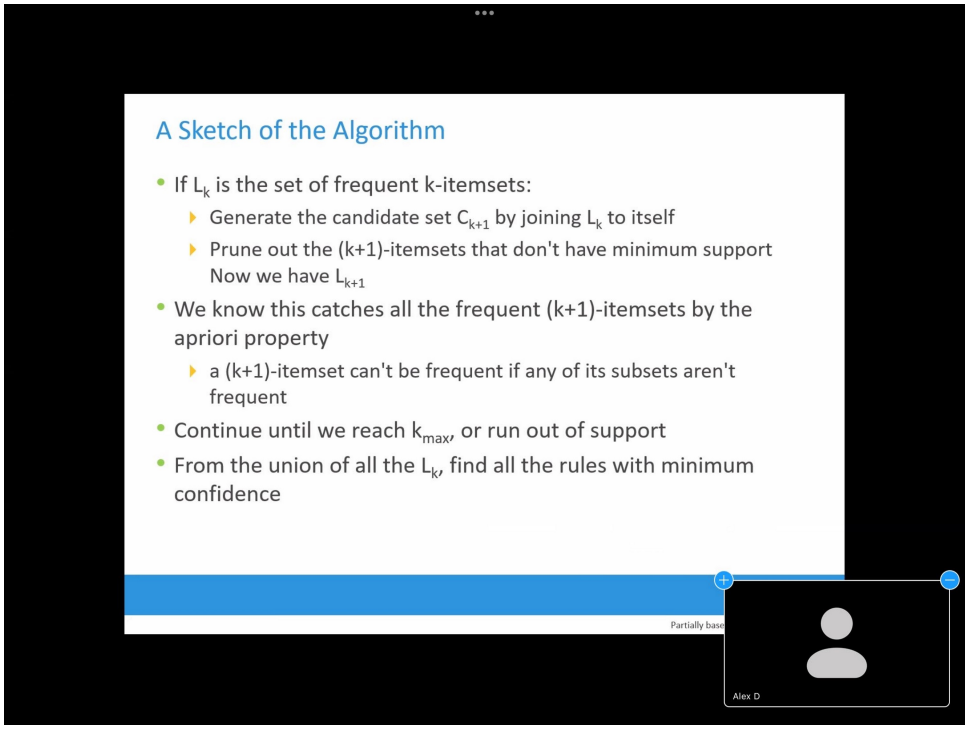
28. **Support** - percent of occurances
29. Superset have at least the same percentage of occurances than set it comes from
30. More importantly - if some set isn't frequent enough than adding elements to it won't help, cause the percentage will fall
31. There is no need to go through all sets. Only through good enough, cause otherwise it'll be computationally hard
32. Forming available rules with minimum **confidence**(percentage of such occurances )
33. But how to exclude consicuencies?

34. 


35. If Lift is large than there is something in it
36. Quite similar to the formula of probability of joint events

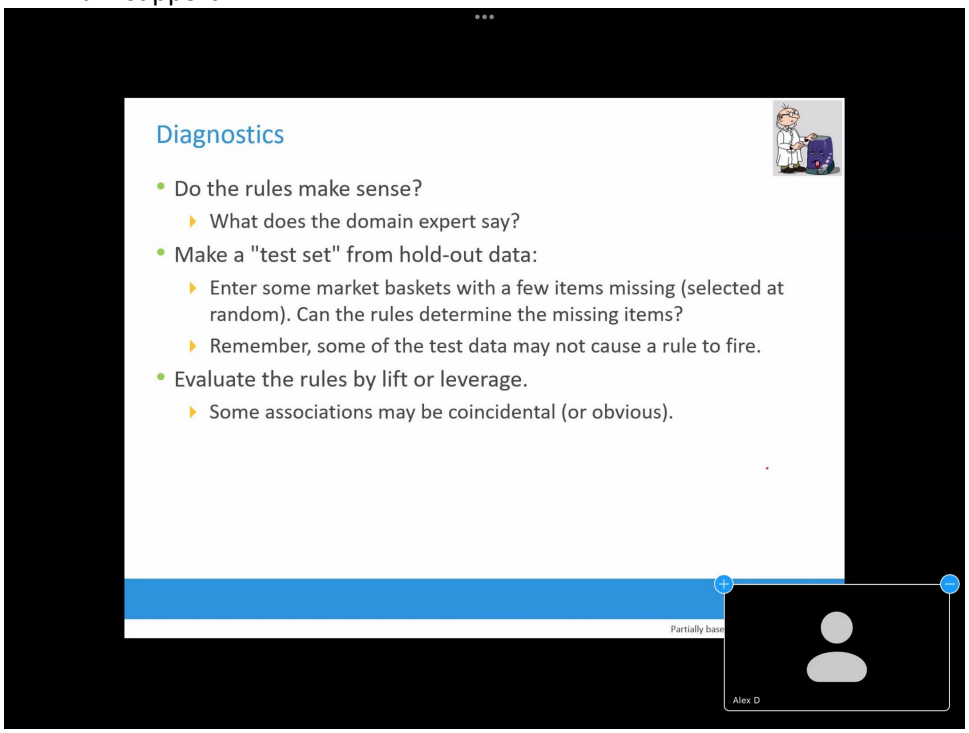
37. 

38.

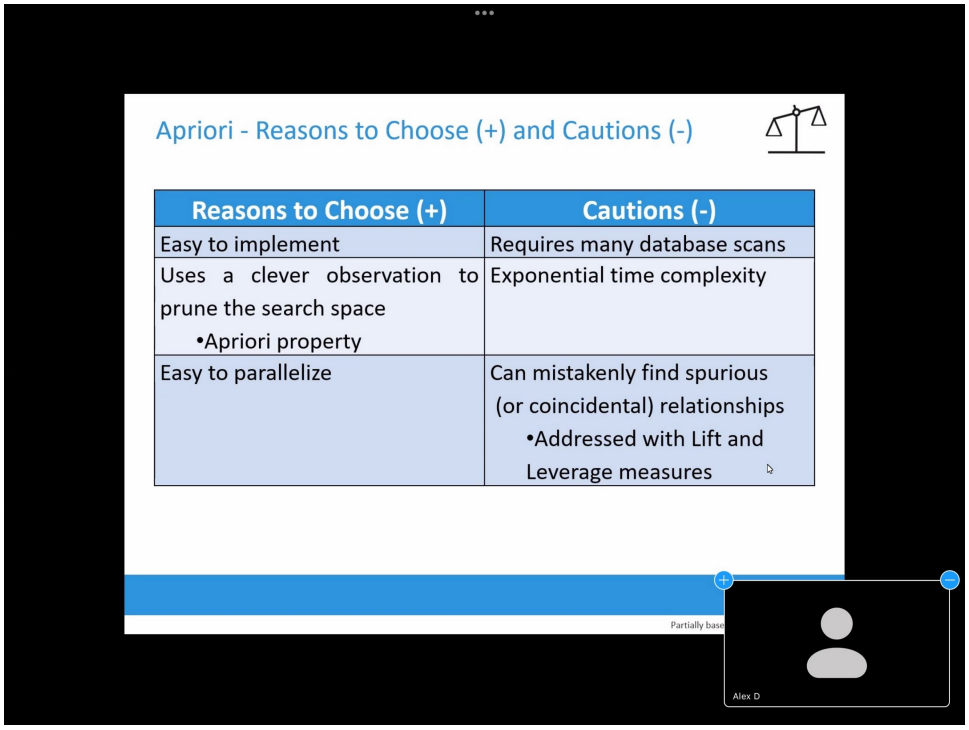



39. Наращиваем порядок наборов пока не дойдём до порядка, на котором нет наборов с minimum support

40.



41.


Apriori - Reasons to Choose (+) and Cautions (-)
Reasons to Choose (+)
Cautions (-)
Easy to implement
Requires many database scans
Uses a clever observation to prune the search space
•Apriori property
Exponential time complexity
Easy to parallelize
Can mistakenly find spurious (or coincidental) relationships
•Addressed with Lift and Leverage measures
Partially base
Alex D