

=========================================

Consider the below dataset of 16 rows,

Google , Facebook – 6 rows each,

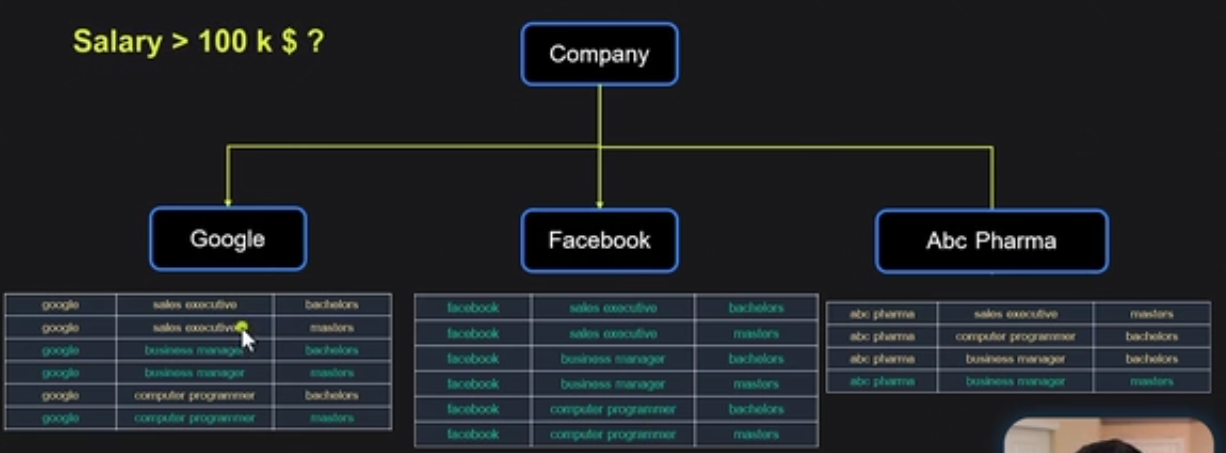
abc pharma – 4 rows

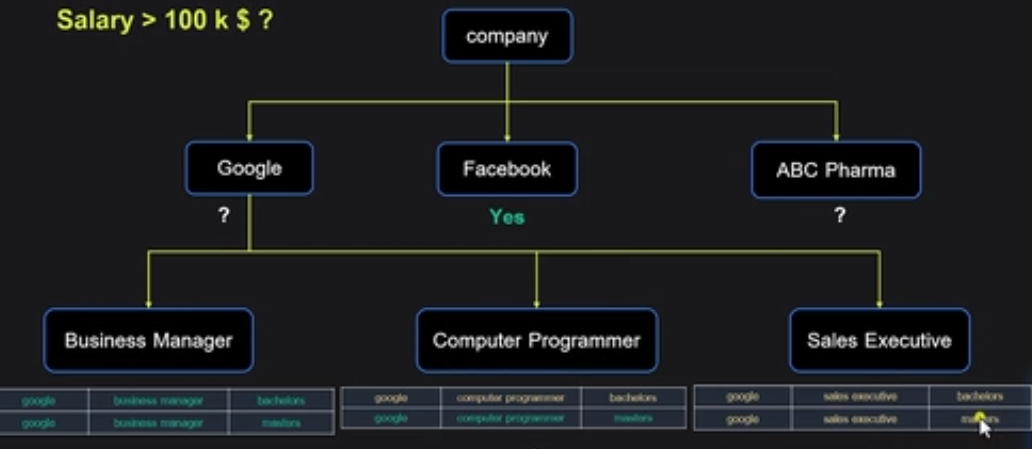
-------------------------

Green rows are salary>100k

Yellow rows are salary<100k



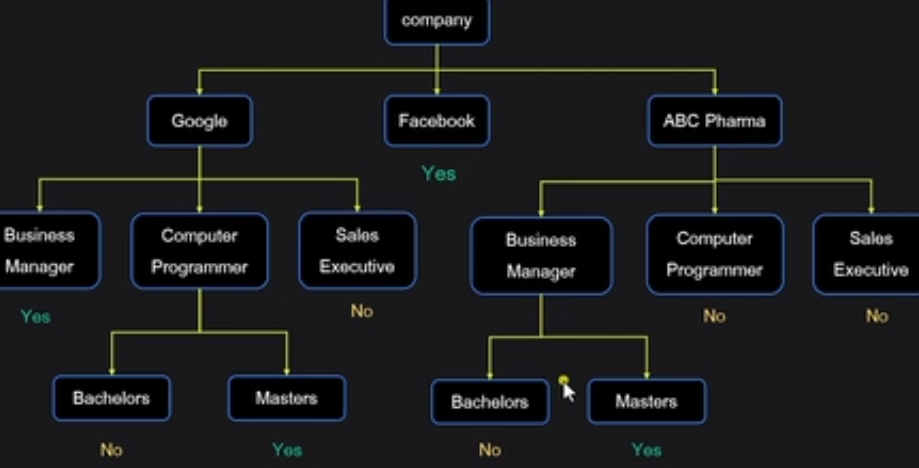




Business Mgr >100k

Computer programmer salary depends on dgree

Sales executive <100k



We could split with respect to the Role title, degree instead of company first.

**How to determine with which feature, we split first?**

We pick a factor such that the number of splits can be minimized.

**=======================================================================**

**Options to select the first split**

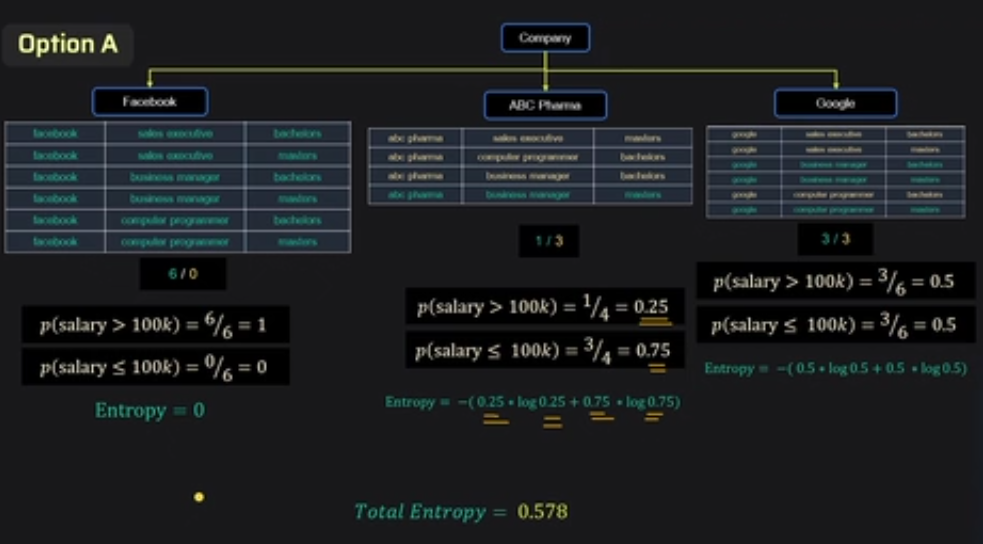
**=======================================================================**

Option A: Company -- Entropy 0.578

Option B: Degree -- Entropy 0.906

Option A is selected over Option B due to its low Entropy, more information Gain

We try all permutations and combinations of splits and select the one with low impurity or Gini index.



**Note:** Total Entropy is the weighted sum



