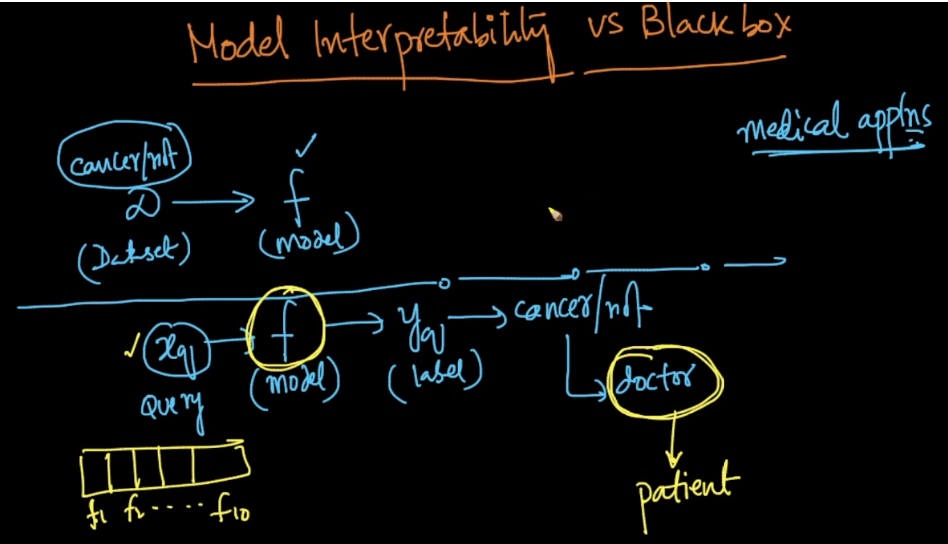
**Black Box Models:**

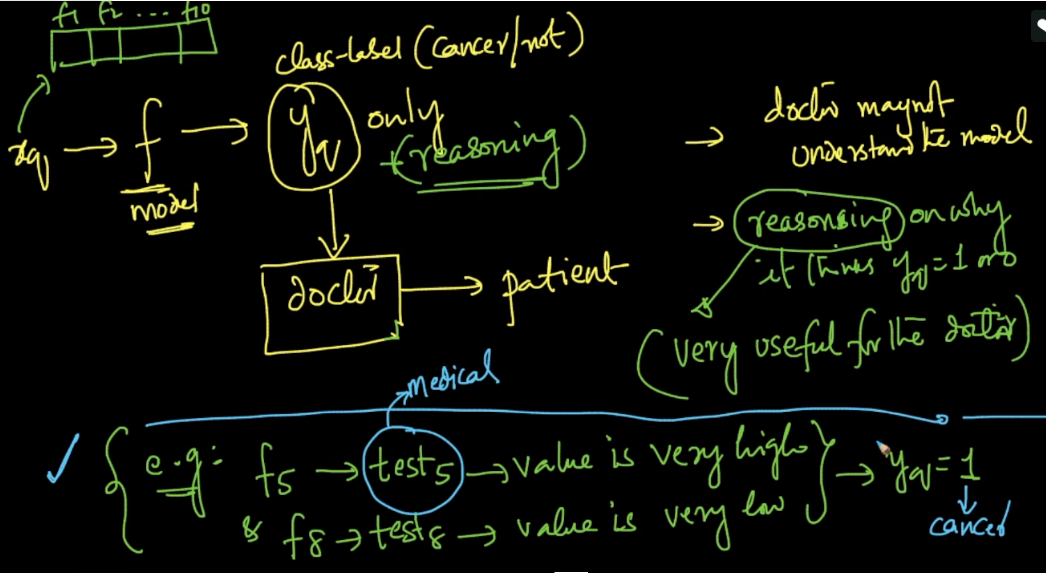
Those models which gives result only as label. For example for problem of cancer detection if it’s output only whether particular patient has cancer or not without describing why it arrives to it’s conclusion is called black box models.

**Model Interpretability:**

Those models which describes why they arrive to a particular conclusion or gives reasoning for the output it is giving is called model interpretability.

Like for cancer detection it says that there is cancer because of particular test performed has high value.



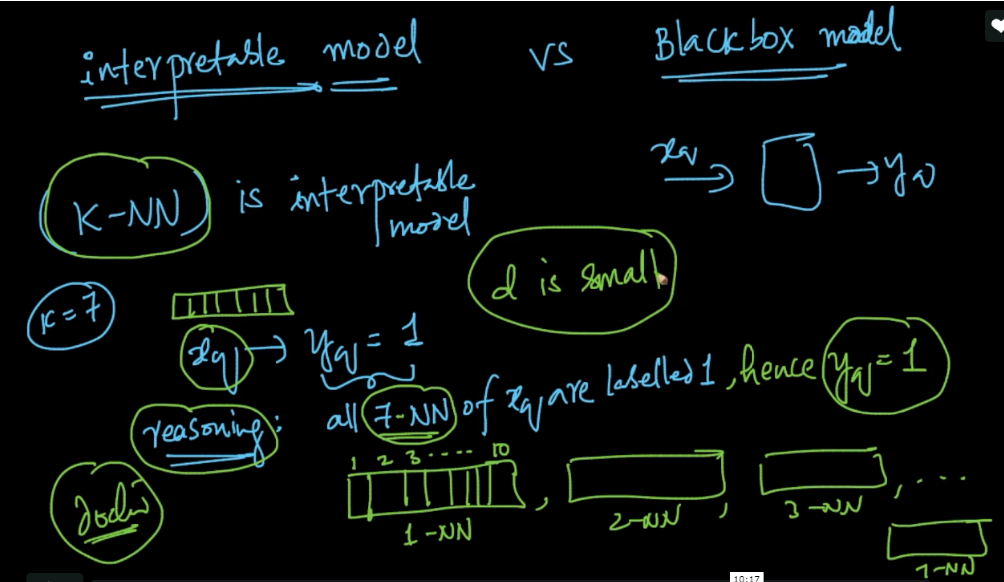


K-NN is interpretable model only if d and k are small. Where d is no of dimensions.

Why?

Because let’s say for cancer detection problem it gives result as that for k=7 with d=10, Majority of NN(patients) have caner with outputting value of each NN(patient).

So each NN or patient has 10 values of test, And doctor can check for 7 similar patients report to give final result for new patient. But if k or d(10k) increases then it becomes almost impossible for Dr. to do cross verification.



**Comments:**

White-box models are the one who can clearly explain how they behave, how they produce predictions and what the influencing variables are. There are two key elements that make a model white-box: features have to be understandable, and the ML process has to be transparent. These models include linear and decision/regression tree models. On the other hand, **black-box models**, such as deep-learning (deep neural network), boosting and random forest models, are highly non-linear by nature and are harder to explain in general. With black-box models, users can only observe the input-output relationship. For example, input the customer profile then output customer churn propensity score. But the underlying reasons or processes to produce the output are not available. Black-box models often result in 1pc to 3pc better accuracy than white-box models, but you sacrifice transparency and accountability. Further reading:<https://datascience.stackexchange.com/questions/22335/why-are-machine-learning-models-called-black-boxes>

* What should be ideal value for d

The optimal 'd' value is problem and domain specific.   
Sometimes, the problem we are solving might require fixed number of dimensions. For example, if we are working on a problem with a dataset having 3000(say) dimensions, then if the domain expert chooses some dimensions and asks only to use those dimensions, then we need to select the subset of features.  
In case, if we are working towards best model performance, then we keep looking for the combination of those features that yield best model performance.