Classification: output is a discrete output.
Regression: output is a point in a spectrum.

Annotator Agreement Metric.
Multiple Annotators.
Annotator agreement is not a dataset quality but bare minimum / start for dataset quality.

Choosing the best algorithm doesn't always mean the most efficient algorithm.

There must be trade off between bias and variance

Some Basic Algo in SUPERVISED ML:

1. K-nearest neighbour : U take test ka data point and compare it to training set and find 5 nearest and closest neighbours. (ORDER = 1)
2. Decision Tree: Basically you go through the features on the basis of their usefulness and you keep dividing the set.
3. Bayesian Classifiers: Naive bayes classifiers.

So far we are assuming the linearity condition. I.e all the features will fall on a straight line. Sometimes this is not going to be correct therefore a classifier which is able to solve such problem

4. Neural network
5. Support Vector Machines: Splits the data and figures out a way to split it non linearly. Though it is not as complex it can only treat binary datasets.
6. Deep neural networks: Can split ‘n’ labels.

UNSUPERVISED ML: Find internal pattern without any labels pre given

1. Clustering: breaks the data in the number of clusters given by us.
2. Parameters: Users need to input a guess
3. Bayesian Parametric Clustering Algo: methods to find similarity in the data. Like Indian buffets or Chinese restaurants. The question is should data stay here or move to another cluster ahead.
4. Topic Model: Statistical model, Finds out the statistical distribution of all the words in all the paper and finds out the words overlapping in all the words over all the paper. Useful fall surveillance.

   Link to coding in class:
   https://colab.research.google.com/drive/1PTQZfaiQvHr_Mpz60jpQ54ByZTwryP9J?usp=sharing