

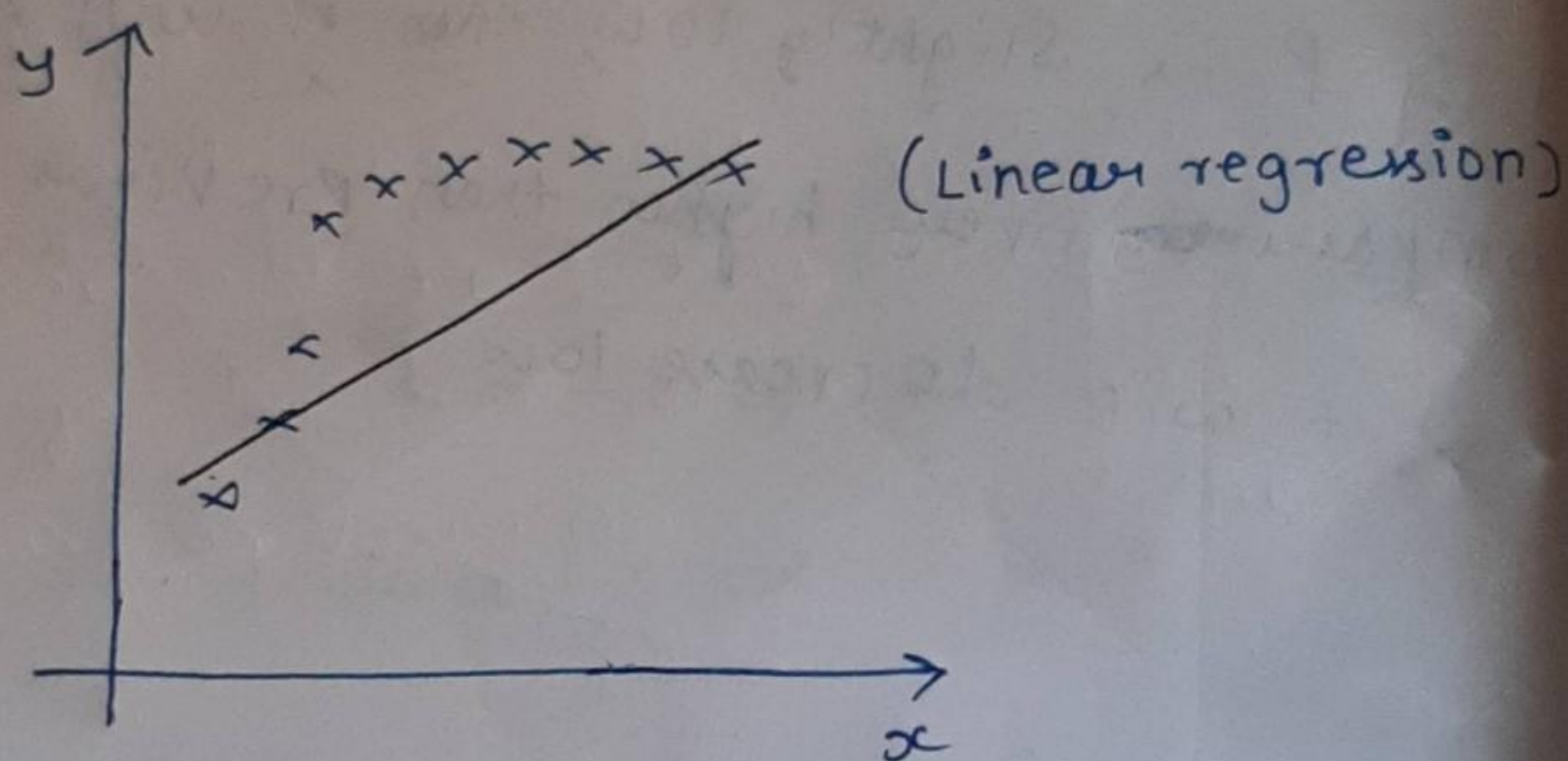
Bias and Variance:-

~~Bias~~

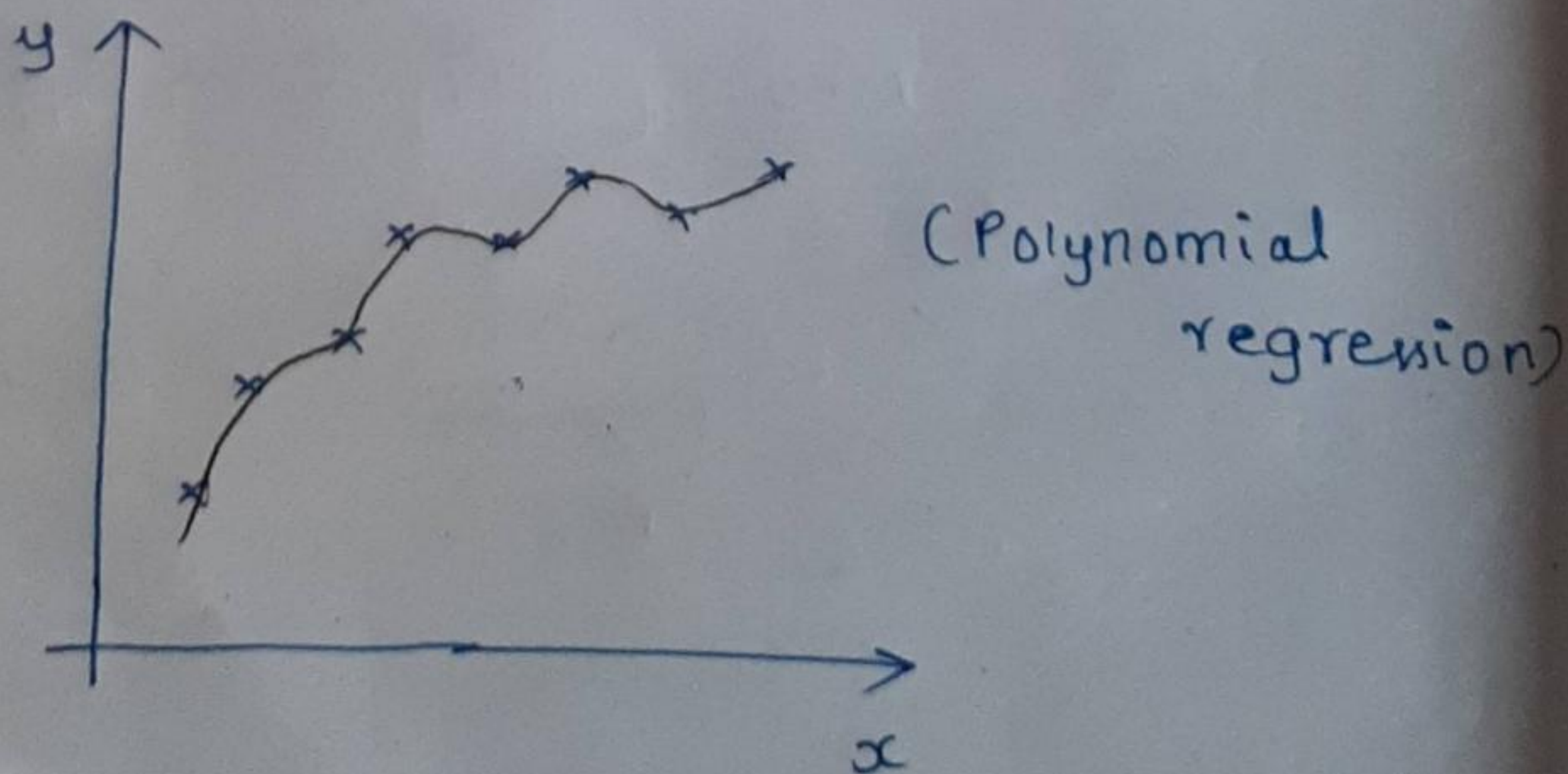
~~overfitting~~

* The overfitting as well as the underfitting means how the predicted data fits the training data as well as testing data.

* To see the explanation, I will use the types of regressions like Linear, Polynomial.



* Basically it is the underfitting model because the distance between the data points and the point on the best fit line is very high.



* Basically it is the overfitting model because the best fit line perfectly fits the data points.

Bias:-

* The inability for a machine learning method like linear regression to capture the true relationship between dependent and independent variable is called bias.

~~* Let's take an example.~~

* It can also be said as error.

Variance:-

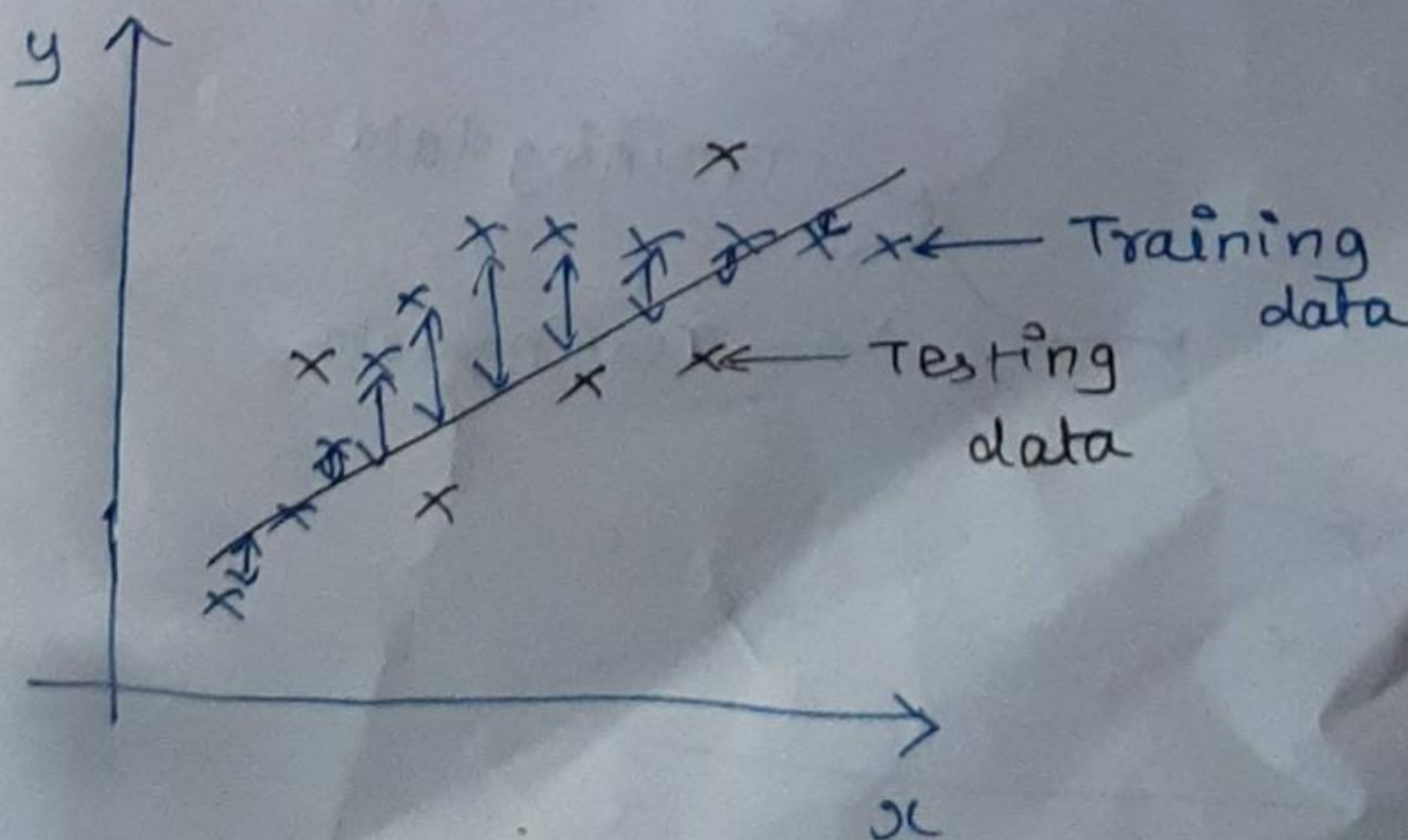
* Variance is the amount that the estimate of the target function will change given different training data or data pts.

* Or it can be said as the difference in fits between data pts is called Variance.

* The variability of model prediction for a given data point which tells us spread of our data is called the Variance of model.

Examples:-

Let's consider a underfitting model:-



* In the above linear regression model we can clearly see that difference between the distance of data points and the points on the predicted line is huge.

* In simple terms we say this as an error, that is what the bias is, underfitting model has the high bias

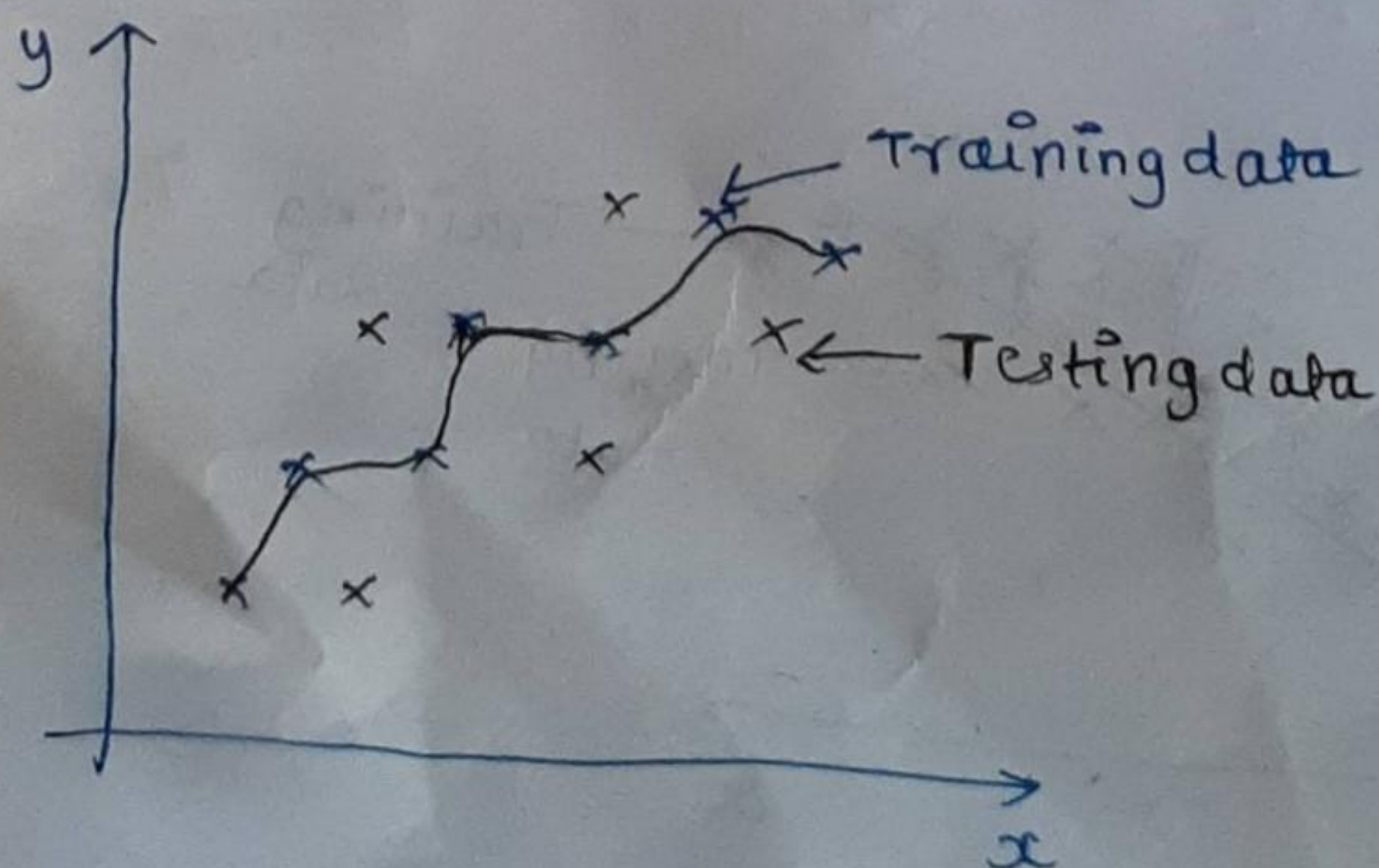
* The bias deals with the training data.

* In the graph we can able to see the testing data pts, The distance b/w points on the predicted line and the testing data pts are not ~~very~~ huge, they are small distance, so this difference or variation is called the Variance

* underfitting model has low Variance.

* Some example algorithms which has the above conditions are Linear regression, Linear discriminant analysis, Logistic regression.

Lets consider a overfitting model:-



* This can be a polynomial regression with degree > 2 . we can clearly see that difference between the distance of data pts and the points on the predicted line is very less.

* So the overfitting model has very low error or low bias

* In the graph we can able to see the testing data pts, the distance b/w the pts and the points on the predicted curve is very high, it is because the model fits the 99.1% with the training data, if the testing data widely differ from any of the training data, then it means model has high

Variance

* overfitting model has high variance

* Some example algorithms which has the above conditions are decision trees, KNN, SVM.

Bias - Variance Trade off :-

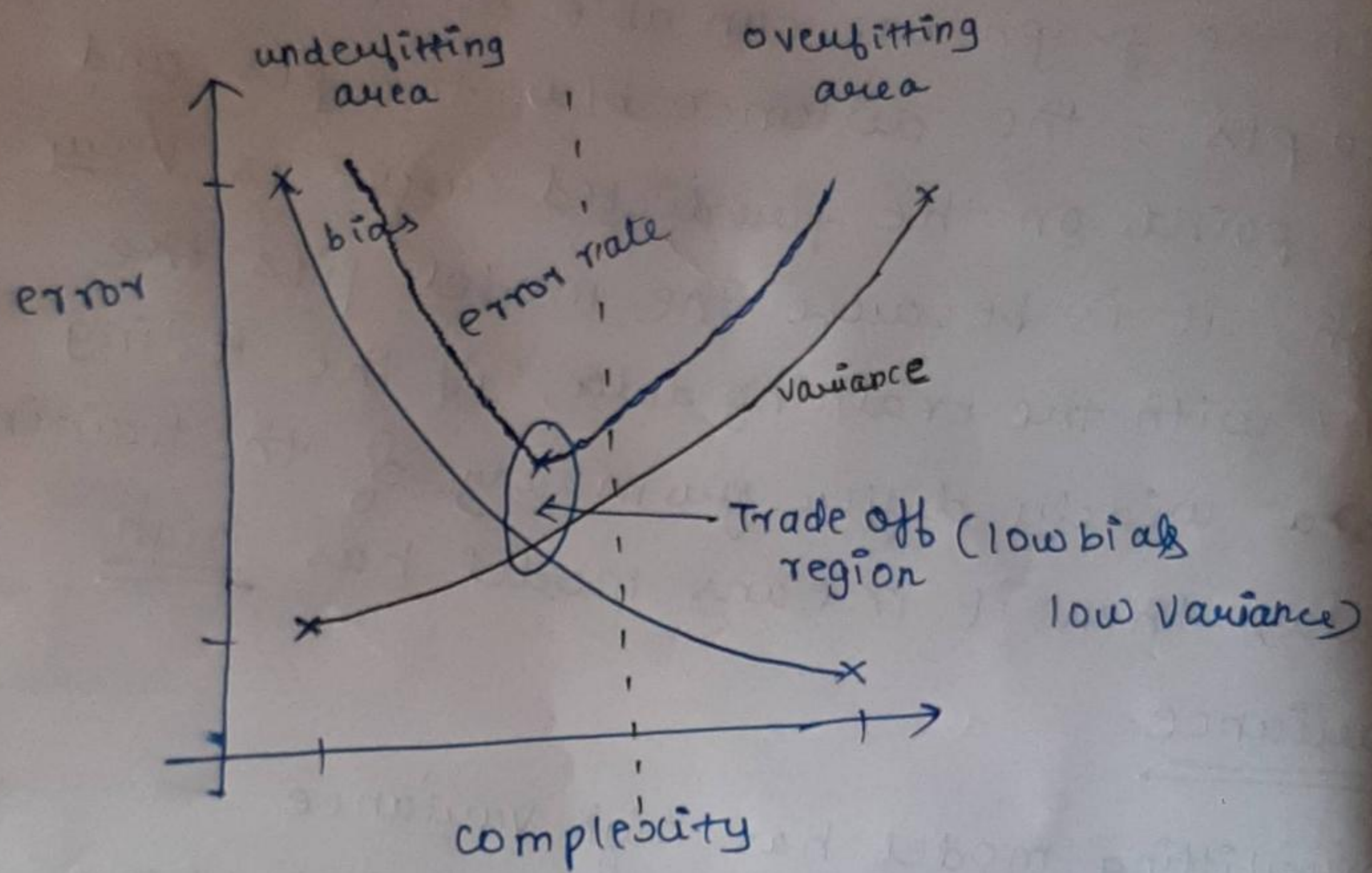
* The goal of any supervised ml algorithm is to achieve low bias and low variance.

~~* The trade off~~

* If algorithm is very simple then it has high bias, low variance, if algorithm is complex then it has low bias, high variance. The condition between these 2 cases were called the trade-off.

* This tradeoff in complexity is why there is a tradeoff between bias and variance

* Simply the algorithm should be more simple and less complex at the same time, literally means model should have low bias, low Variance.



* It is the combination of representing both conditions for underfitting and overfitting.

* The error rate will decrease in underfitting but once, it starts to overfit the model, the variance will increase.

* So we should stop before that, and the generalized region is called the trade-off.