Bias and Variance

- Main goal of supervised learning: **prediction**
- **Prediction error** ~ reducible + irreducible error

Irreducible - reducible error

- **Irreducible:** noise — **don't minimize**
- **Reducible:** error due to unfit model — **minimize**
- **Reducible error** is split into bias and variance

Bias

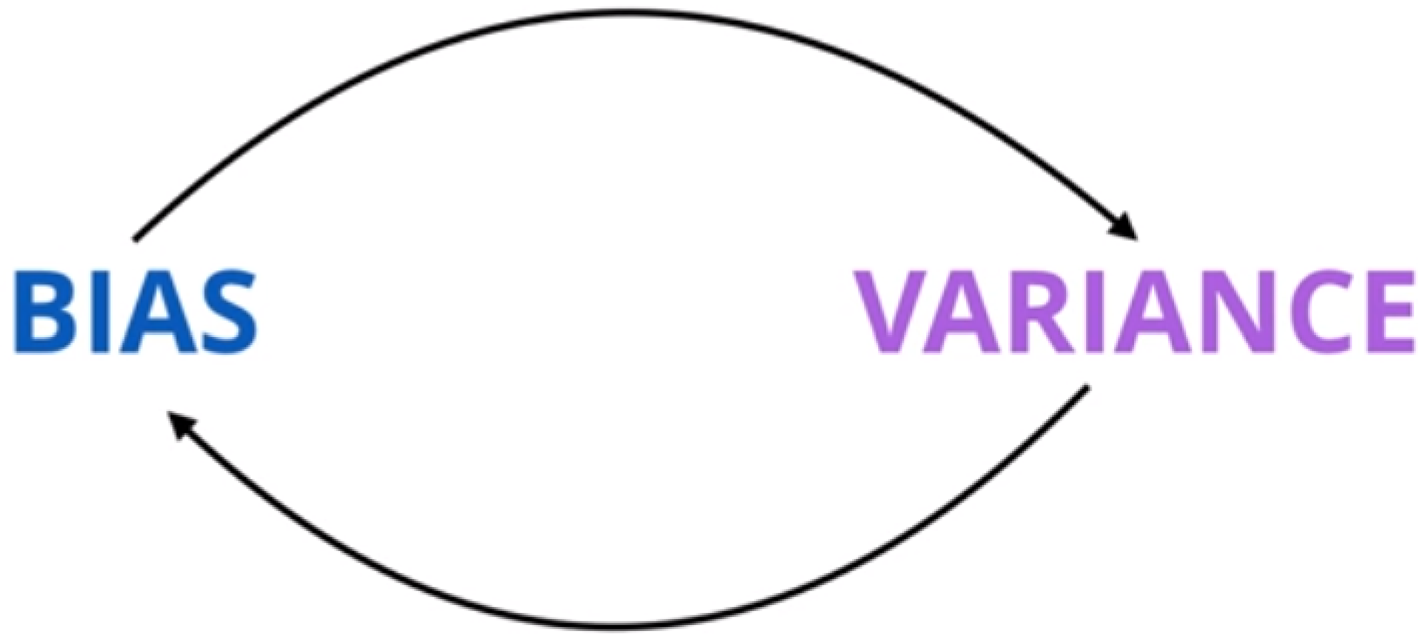
- Error due to bias: wrong **assumptions**
- Difference **predictions** and **truth**
    - using models trained by specific **learning algorithm**

- Complexity of model
- More restrictions lead to high bias

Variance

- Error due to variance: error due to the sampling of the **training set**
- Model with high variance fits **training set** closely

Bias-variance tradeoff



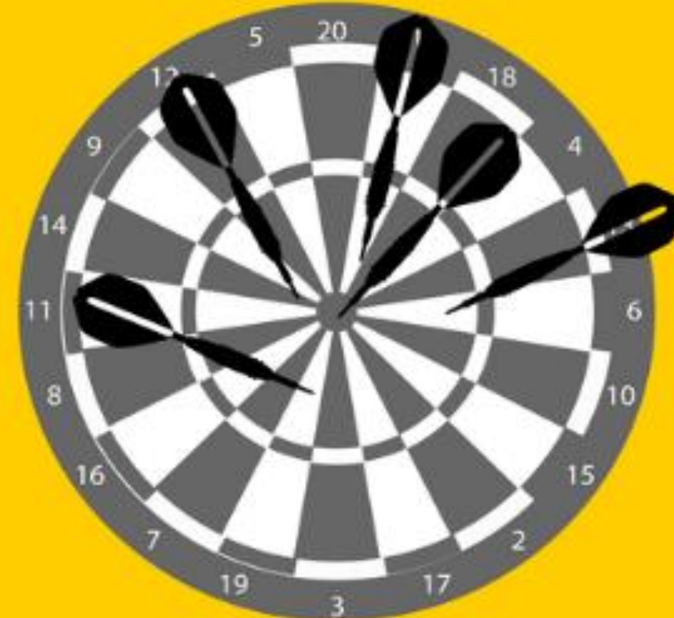
low **bias** - high **variance**
low **variance** - high **bias**

High Bias

Low Variance

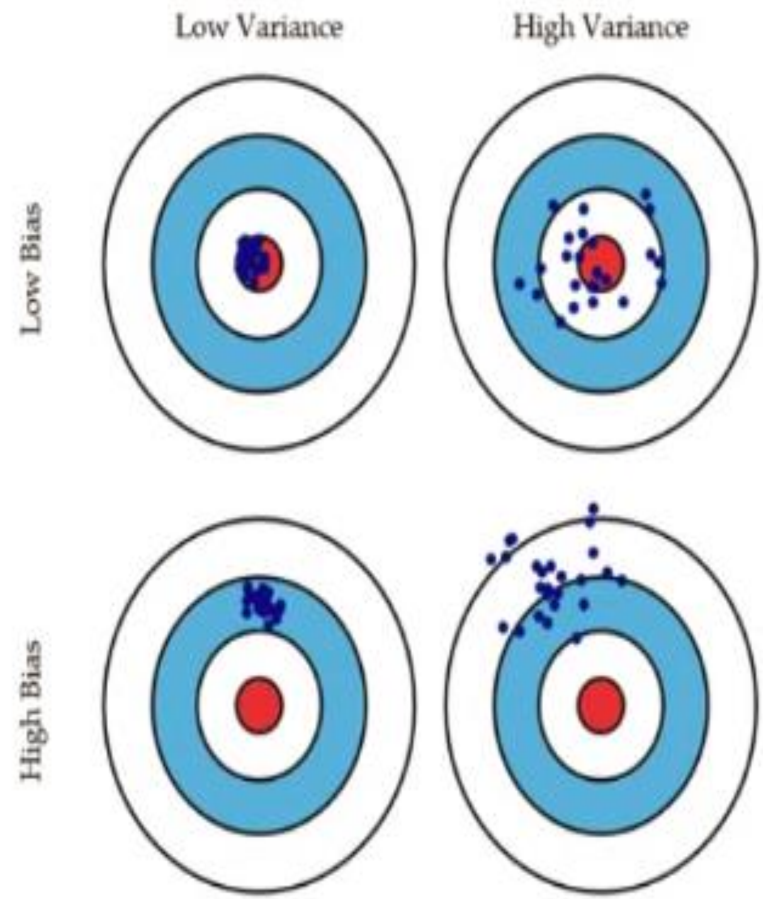
High Variance

Low Bias



High bias, low variance algorithms train models that are consistent, but inaccurate *on average*.

High variance, low bias algorithms train models that are accurate *on average*, but inconsistent.



Let's say we have model which is very accurate, therefore the error of our model will be low, meaning a low bias and low variance as shown in first figure. All the data points fit within the bulls-eye. Similarly we can say that if the variance increases, the spread of our data point increases which results in less accurate prediction. And as the bias increases the error between our predicted value and the observed values increases.

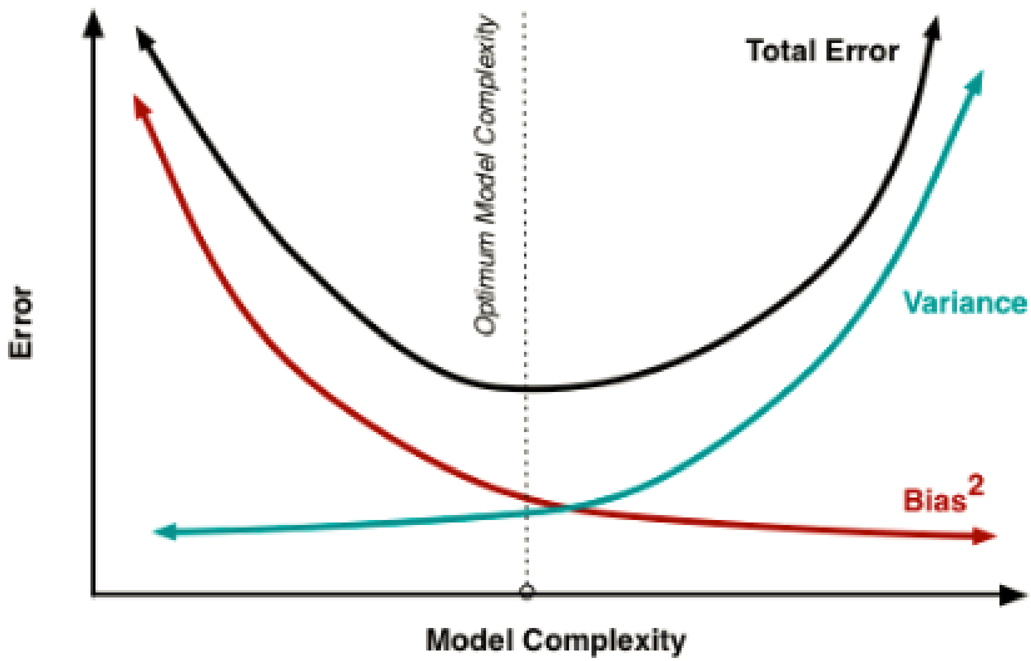
Overfitting

- **Accuracy** will depend on dataset **split** (train/test)
- High variance will **heavily** depend on **split**
- Overfitting = model fits **training set** a lot better than **test set**
- Too specific

Underfitting

- Restricting your model **too much**
- High **bias**
- Too **general**

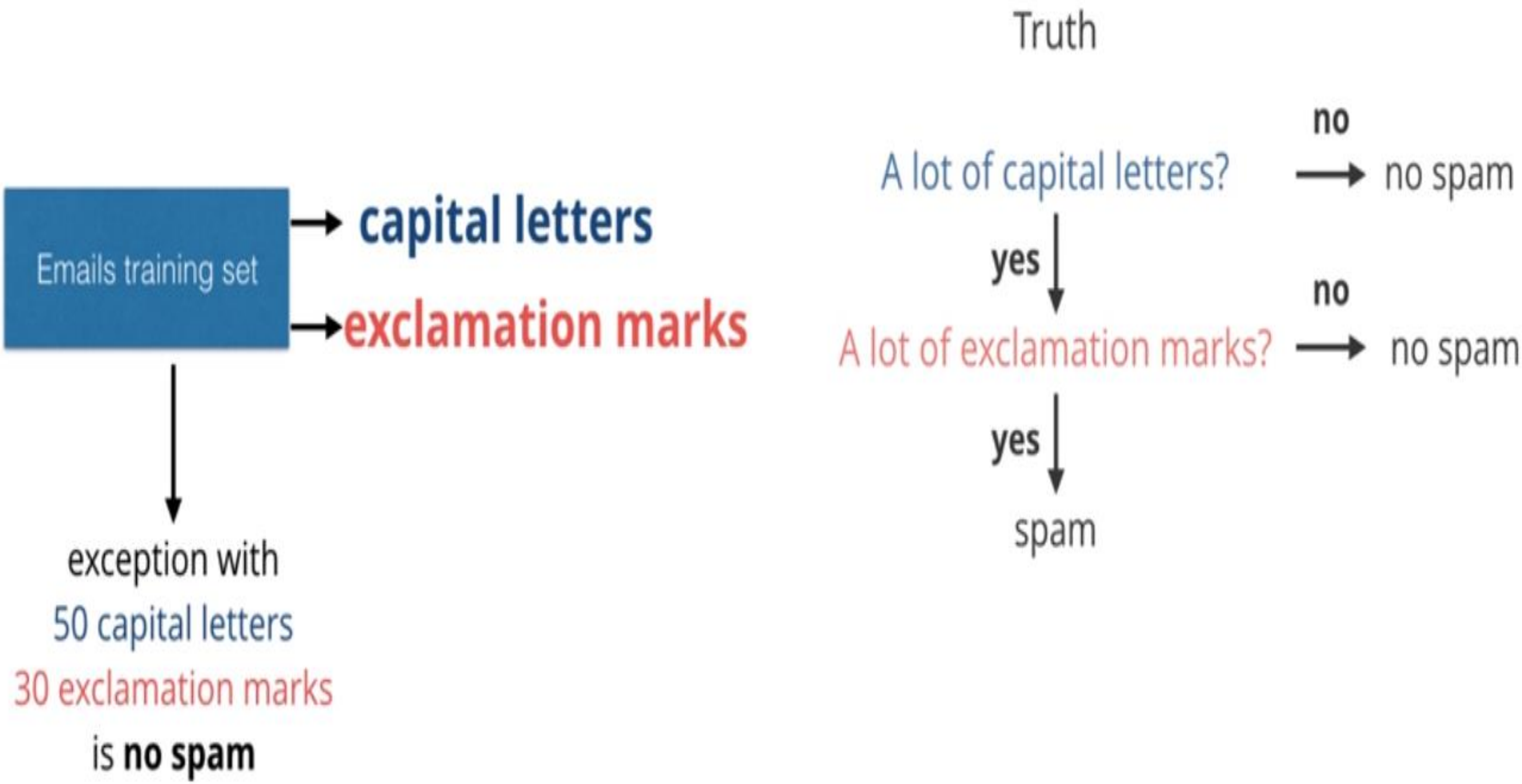
Now how this bias and variance is balanced to have a perfect model? Take a look at the image below and try to understand.



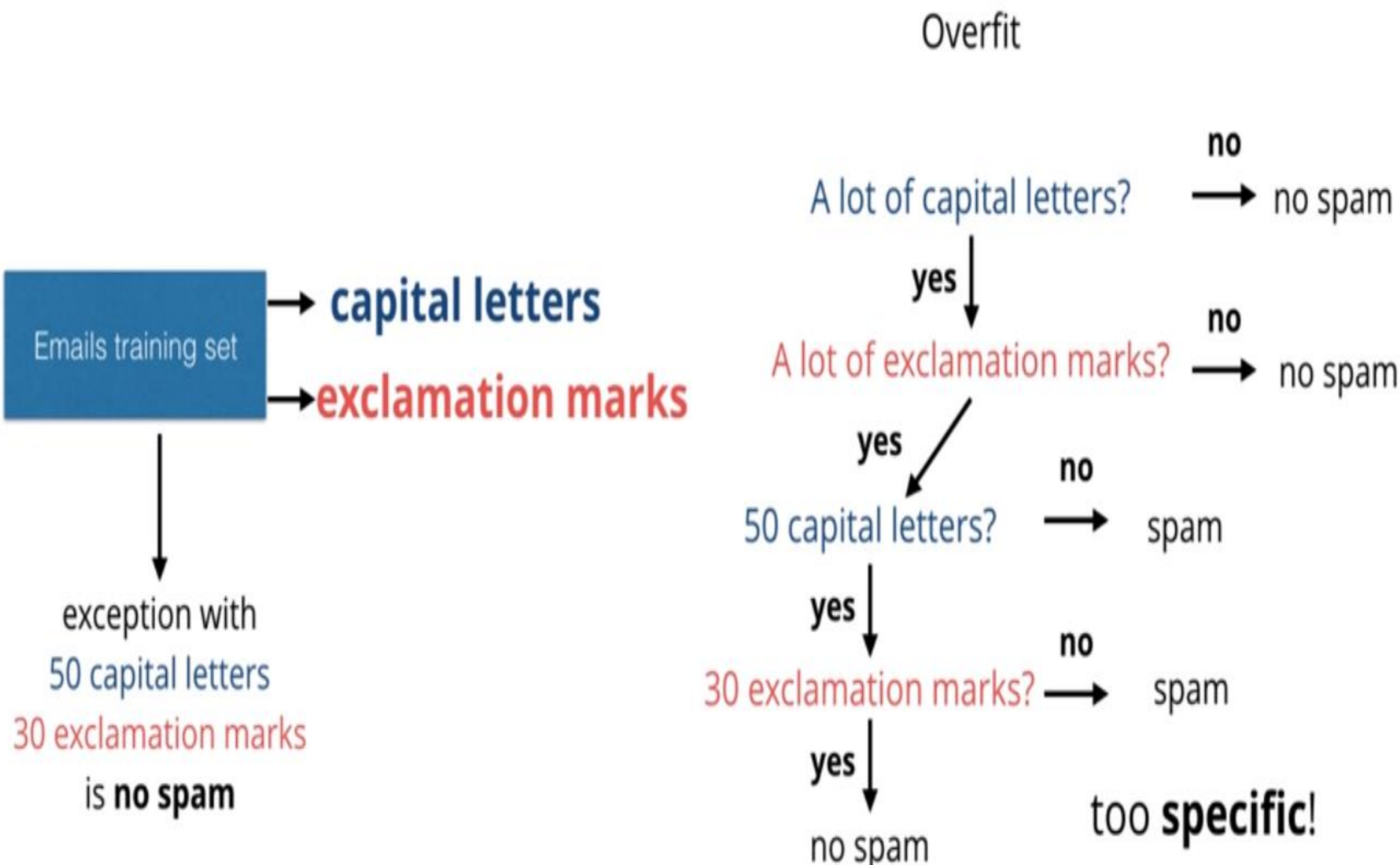
As we add more and more parameters to our model, its complexity increases, which results in increasing variance and decreasing bias, i.e., overfitting. So we need to find out one optimum point in our model where the decrease in bias is equal to increase in variance. In practice, there is no analytical way to find this point. So how to deal with high variance or high bias?

To overcome underfitting or high bias, we can basically add new parameters to our model so that the model complexity increases, and thus reducing high bias.

Example - spam or not?



Example - spam or not?



Example - spam or not?

Emails training set → **capital letters**

Emails training set → **exclamation marks**

Underfit

More than 10 capital letters? $\xrightarrow{\textbf{no}}$ no spam

yes ↓

spam

too **general**!

Now, how can we overcome Overfitting for a regression model?

Basically there are two methods to overcome overfitting,

- Reduce the model complexity
- Regularization